Guidelines and Standards for Evidence Synthesis in Environmental Management

VERSION 5.0

2018

Editors: Andrew S Pullin, Geoff K Frampton,
Barbara Livoreil & Gillian Petrokofsky

Read the guidelines online here

Please cite as: Collaboration for Environmental Evidence. 2018. *Guidelines and Standards for Evidence synthesis in Environmental Management*. Version 5.0 (AS Pullin, GK Frampton, B Livoreil & G Petrokofsky, Eds) www.environmentalevidence.org/information-for-authors. [date of access]

Please note that these guidelines will be periodically updated and each update recorded (see Updates and Corrections). Major updates will be announced through social media.

*** Please help us improve the Guidelines by taking a few minutes to complete this short survey***

Acknowledgements

Thanks to the Guidelines Editorial Group for this version of the CEE Guidelines and Standards (in alphabetical order): Geoff Frampton, Barbara Livoreil, Gillian Petrokofsky and Andrew Pullin

We thank all authors contributing to this version of the CEE Guidelines and Standards and the papers on which some sections were based (in alphabetical order): Helen Bayliss, Alison Bethel, Gary Bilotta, Frédérique Flamerie de Lachapelle, Jacqui Eales, Geoff Frampton, Julie Glanville, Neal Haddaway, Katy James, Christian Kohl, Julia Koricheva, Elena Kulinskaya, Magnus Land,

Barbara Livoreil, Biljana Macura, Gillian Petrokofsky, Andrew Pullin, Nicola Randall, Shannon Robalino, Sini Savilaakso, Jessica Taylor and Wen Zhou.

This version of the CEE Guidelines and Standards has developed from previous versions and we thank all those who have contributed to those versions.

We thank the many managers, policy formers and scientists who have given us constructive feedback on the review process and those who have themselves contributed to CEE Evidence Syntheses.

Previous Versions

Guidelines (v4.2)

Guideline Translations

Thanks to members of our global community, the CEE Guidelines have been translated into other languages. Please note that CEE does not have the resources to endorse these documents as accurate translations and will not be updating them when changes are made to the English guidelines.

We thank the translators for making CEE more accessible to evidence synthesists and users around the world.

Japanese (v5.0 - as of Aug 7, 2020)

Translated by: Ko Konno

Spanish (v4.2)

Translated by: CEE Chile, Ana Benítez-López, Alejandro Martínez- Abrain, Francisca Rodríguez, Adrián Villaseñor

Chinese (v4.2)

Translated by: Dr Jing-Chun Fan

Aims and Scope

Evidence review and synthesis methodology (hereafter referred to as 'evidence synthesis') is now in widespread use in sectors of society where science can inform decision making and has become a recognised standard for accessing, appraising and synthesising scientific information. The need for rigour, objectivity and transparency in reaching conclusions from a body of scientific information is evident in many areas of policy and practice, from clinical medicine to

social justice. Our environment and the way we manage it are no exception and there are many urgent problems for which we need a reliable source of evidence on which to base actions. Many of these actions will be controversial and/or expensive and it is vital that they are informed by the best available evidence and not simply by the assertions or beliefs of special interest groups. For evidence synthesis to be credible, legitimate and reliable, standards regarding its conduct need to be clearly defined. Such standards include examining possible sources of bias both in the evidence and in the way the review and synthesis is conducted. In so doing, the goal is to provide an explicit level of confidence in the findings to the end-user. Here we present the latest guidelines for the planning and conduct of CEE Evidence Syntheses (separated into Systematic Reviews and Systematic Maps see below) in environmental management.

The guidelines and standards for CEE Evidence Syntheses (including the planning and review stages) have been adapted from methodologies developed and established over more than two decades in the health sciences (Higgins & Green 2009), informed by developments in other sectors such as social sciences and education (Gough et al. 2012) and tested through practice in developing the CEE Library of Evidence Syntheses. Through undertaking and peer reviewing CEE Evidence Syntheses, researching and adapting existing methodologies, and through analysis of procedures and outcomes, CEE contributors have developed specific guidelines for application to environmental management and the types of data and study designs that are prevalent in environmental research. Whilst past CEE Systematic Reviews and Systematic Maps may provide some guidance, our advice is not to assume that past practices are sufficient for future CEE Standards. This document refers to examples of best practice and CEE is constantly trying to improve standards of evidence review and synthesis.

Although the basic ethos of evidence synthesis is generic, environmental methodologies are often different in nature and application from those in other fields and this is reflected in these guidelines. At first glance, many of the approaches may seem routine and common sense, but the rigour and objectivity applied at key stages, and the underlying philosophy of transparency and independence, sets them apart from the majority of traditional reviews published in the field of applied ecology (Roberts et al. 2006, O'Leary et al. 2016). Evidence syntheses are being commissioned by a wide range of organisations in the environmental sector and the need for common guidelines and standards, and collaborative development of the methodology, is critical to the formation of an openly accessible and credible evidence base that functions as a public good. We argue that, once more widely established, CEE methodology will significantly improve the identification and provision of evidence to inform practice and policy in environmental management. For this methodology to have an impact on effectiveness of our actions, more environmental scientists and other stakeholders need to get involved in the conduct of CEE Evidence Syntheses. For those intending to conduct evidence reviews syntheses, these guidelines are provided in the spirit of collaboration and we encourage you to contribute your work to the CEE, use and improve these guidelines, and help establish an evidence-based framework for our discipline.

Who are these guidelines and standards for?

These guidelines are primarily aimed at those teams intending to conduct a CEE Evidence Synthesis. The structure of the document takes the reader through the key stages from first

consideration of the need for an evidence synthesis to the dissemination of the findings. Novice Review Teams should not expect that these guidelines alone will be sufficient support to conduct an evidence synthesis to CEE standards. They are guidelines only and do not replace formal training in CEE methodology.

We hope that these guidelines and standards will also be of use to those considering commissioning the conduct or using the findings of a CEE Evidence Synthesis and for stakeholders who may become involved in their planning. In this context the Guidelines provide standards for conducting and reporting syntheses that commissioners and stakeholders can expect to be demonstrated by their authors.

Finally, these guidelines set a standard for the conduct of evidence syntheses and are therefore relevant for decision makers using evidence from CEE and wishing to understand the nature of the CEE process and how it provides a reliable assessment of the evidence.

Some basics

For clarity of process, the guidelines are split into separate sections. There is obviously considerable overlap between planning, conducting and reporting and we cross reference as much as possible to avoid undue repetition but some is unavoidable. We use examples of completed Systematic Reviews and Systematic Maps in the CEE library to illustrate each stage of the process and to highlight key issues. A glossary is provided on the CEE website but here are some key definitions.

- **What is Evidence?** The Oxford English Dictionary definition is: "The available body of facts or information indicating whether a belief or proposition is true or valid." Note: The degree to which any information constitutes evidence depends on the question being asked and the context in which it is asked. For the purposes of these guidelines evidence is assumed to have been generated by scientific studies, which are referred to as "primary" research. The assimilation and combination of this evidence using the synthesis methods we describe here (Systematic Reviewing or Systematic Mapping) is referred to as "secondary" research.
- **What is a Systematic Review?** A Systematic Review is an evidence synthesis method that aims to answer a specific question as precisely as possible in an unbiased way. The method collates, critically appraises, and synthesizes all available evidence relevant to the question. Reviewers use pre-defined methods to identify risks of bias in the evidence itself, and to minimise bias in the way evidence is identified and selected, and thus provide reliable findings that could inform decision making.
- **What is a Systematic Map?** A Systematic Map is an evidence synthesis method that aims to provide an accurate description of the evidence base relating to a particular question. The method collates, codes, and configures all available evidence relevant to the question. Reviewers use pre-defined methods to minimize bias in the way the evidence is identified and selected. A descriptive overview of the evidence base is developed that could inform further research and synthesis (e.g. by revealing knowledge gaps and identifying more specific questions suitable for Systematic Review).

The differences between Systematic Reviews and Systematic Maps, and guidance on how to decide whether to conduct a Systematic Review or a Map, are explained in more detail in the following sections. In essence, both approaches review evidence and start out in the same way, being protocol-based and requiring systematic searches and systematic evidence selection techniques, but their mode of synthesis, analyses and outputs differ. A Systematic Review provides an aggregate answer to a specific question, whereas the output of a Systematic Map is a configurative, descriptive characterisation of the evidence base. Systematic Reviews may be confirmatory and hypothesis-testing, whereas Systematic Maps may be more exploratory and hypothesis-generating, although this is not a rigid distinction.

Section 1

Process Summary: Registration, Publication and Dissemination of a CEE Evidence Synthesis

Last updated: September 27th, 2017

This section provides a summary of the steps in the conduct of a CEE Evidence Synthesis (Systematic Review or Systematic Map), an overview of how authors register their Evidence Synthesis with CEE and of the process of submission and peer review that ensures CEE Evidence Syntheses are conducted to high standards.

1.1 The CEE registration, submission and deposition process

CEE operates an open-access policy and all of its contributors' Protocols, Systematic Reviews and Systematic Maps are published (subject to peer review) in its open-access journal *Environmental Evidence*(www.environmentalevidencejournal.org). This provides authors with a high level of visibility for their publications.

Here we set out the process for registering intent to conduct and contribute a CEE Evidence Synthesis, and for publishing Protocols and Final Reports in *Environmental Evidence*. High standards of reporting are expected on the conduct of a CEE Evidence Synthesis and this starts with the submission of a Protocol and continues through to the provision of supplementary material such as data extraction spreadsheets and a list of excluded articles. Full instructions for authors on preparation of manuscripts, including templates and checklists, are available from the *Environmental Evidence* journal website at http://environmentalevidencejournal.biomedcentral.com/submission-guidelines

In cases below where the guidance applies equally to a Systematic Review or Map we refer to these collectively as "Evidence Syntheses". Registration and submission of an Evidence Synthesis to Environmental Evidence is an interactive stepwise process as follows;

1. Draft Protocols are submitted to *Environmental Evidence* through its electronic submission system (www.environmentalevidencejournal.org). This serves as an application for registration to conduct a CEE Evidence Synthesis. The draft Protocol will

be sent out for peer review. Comments will be returned to the authors and appropriate revisions may be requested to finalise the Protocol. By publishing your Protocol in *Environmental Evidence*you are registering with CEE your intent to conduct, and submit to this journal for publication, a CEE Evidence Synthesis. You will be asked to confirm that you and your co-authors are aware of and agree with this commitment when you submit by agreeing to the following statement – 'The authors hereby submit our Protocol for publication in *Environmental Evidence*. By doing so we register with CEE our intent to conduct and submit to this journal a full and original Systematic Review/Map Report for publication and archiving in the CEE Library'.

2. The finalised Protocol is published in *Environmental Evidence*, posted on the CEE website and the Evidence Synthesis is then formally registered as being 'in progress'. At this point a dedicated review webpage will be created on the CEE website and can be used by the authors to post updates and news.
3. Once conducted and written up, submission of a draft Evidence Synthesis manuscript to *Environmental Evidence*follows the same electronic submission process. If acceptable after an initial screening, the draft Evidence Synthesis will be sent out for peer review. Comments will be returned to the authors and appropriate revisions may be requested before acceptance.
4. The revised and completed Evidence Synthesis (and its associated supplementary material) will be published in *Environmental Evidence* and posted as finalised on the website in the CEE Library.

Please note that CEE does not accept manuscripts of unregistered Evidence Syntheses (i.e. those without a previously registered and published Protocol) nor does it accept retrospective Protocols or registration of already completed Evidence Syntheses. CEE reserves the right to reject Protocols and Evidence Syntheses if they do not meet our standards or are otherwise inappropriate.

Article Processing Charges for both the Protocol and final report are payable (http://environmentalevidencejournal.biomedcentral.com/about) in line with most open-access journals. Protocol APCs include a charge for CEE Registration.

CEE operates a supportive policy for review teams undertaking Evidence Syntheses and seeks to provide help and guidance, particularly during the Protocol finalisation stage (including through web-based support materials and training events) to increase the chances of Systematic Reviews and Systematic Maps being successfully completed.

1.2 Supplementary materials

The transparency of Evidence Syntheses is enhanced by the provision of a range of mandatory supplementary materials. Some can be provided as appendices whilst others may be posted as additional files on the review webpage. For a full checklist see Section 10.

1.3 Further dissemination of findings

Although CEE Evidence Syntheses are designed to be reliable sources of evidence they do not necessarily make the evidence very accessible to a non-scientific readership. After all the work of searching, screening, appraising, extracting and synthesising evidence and writing the report, it is worth considering whether the full evidence synthesis format is sufficient or appropriate for disseminating the key outcomes to your target audience. The publication of the full CEE Evidence Synthesis constitutes an important resource and a transparent audit trail of methodology but may not be suitable as a dissemination tool to reach decision makers. Other formats such as policy briefs, executive summaries and guidance notes can be developed and posted on the Evidence Synthesis webpage (as well as being disseminated elsewhere). Such documents often require some special skills in order to make the conclusions and recommendations, as well as their justification, accessible to a non-scientific audience. They can be written by the review team, but can also be designed by a specialist or during meetings with policy makers and/or practitioners and managers.

1.4 Updating an Evidence Synthesis

Evidence syntheses can only be accurate assessments of the evidence base when they are up to date. As soon as the search is completed the reliability of an evidence synthesis as a synthesis of 'all available evidence' begins to decline. The rate of decline is dependent on the rate of publication of new studies and so varies from subject to subject. An outdated Systematic Review or Systematic Map may be misleading, so they should periodically be updated. Fortunately the process of updating a CEE Evidence Synthesis should not be as burdensome as the original process, provided that accurate reporting was achieved and good records were kept of the original process. We encourage the publication and archiving of as full a record as possible of all procedures and outcomes as supplementary materials. At the time of writing, updating a CEE Evidence Synthesis is yet to be completed; we suggest considering updating a Systematic Review or Systematic Map 3-5 years after publication depending on the rate of publication of new primary studies. The process for registering an update is the same as for an original Evidence Synthesis and should begin with an updated Protocol. Updates can be proposed by original authors, other review teams or a combination and should be justified in terms of new studies potentially strengthening the evidence base or the potential to improve the synthesis in some way.

Section 2

Identifying the need for evidence, determining the Evidence Synthesis type, and establishing a Review Team

Last updated: September 2nd 2020

2.1 Determining the need for evidence

In trying to find solutions to problems, the cause of a change, or to decide among alternative interventions to achieve desired outcomes, individuals or groups may identify a need for evidence. This chapter provides generic guidance on how to identify evidence needs to inform decision making. In doing so we provide some guidance on the initial steps that may, or may not, result in the planning and commissioning of a CEE Evidence Synthesis. It is not our intention here to describe the policy process or how management decisions are made.

The need for evidence relating to a question of concern in policy or practice can arise in many ways, ranging from the scientific curiosity of individual researchers to global policy development. Identifying and agreeing priority issues and the need for evidence to inform decisions are often iterative processes involving dialogue among many different individuals and organisations. In the process of deciding how to spend limited resources to achieve organisational objectives, there is an opportunity for that decision to be informed by the best available evidence. Identifying exactly what evidence would help decision-making is therefore worth some thought and discussion.

The question addressed by a CEE Evidence Synthesis often arises from a concern, a challenge, a conflict or a decision that needs to be taken, for which the decision-makers would like to be informed by the best available evidence. Often, decision-makers would like to know how to intervene to solve a problem or what evidence is available to help them make the best decision. They may like to know whether evidence has accumulated, studies or trials have been conducted, or effects have been measured.

Questions arise in many forms and examples of common generic question types are listed in Table 2.1. Examples of scenarios that might generate a need for evidence are shown in Box 2.1.

Table 2.1. Common types of policy problems and concerns in Environmental Management

Answer being sought	Example question
Greater understanding or predictive power	What is the role of biodiversity in maintaining specific ecosystem functions (e.g. biogeochemical cycles)? Here, a specific problem may be assessed to know whether it is really a problem and, if so, how big it is, and what are the significant drivers of changes.
Impacts of exposure to anthropogenic stressors	What is the impact of wind farm installations on bird populations? This type of request often addresses the effect of an exposure to a device, management practice or other stressor (e.g. pollutant) on biodiversity.
Socio-economic outcomes	What are the anticipated costs of the impacts of invasive species on health or agriculture? This type of request may require datasets collected by economists and social scientists, and their associated specific analytical tools.

Intervention effectiveness	How effective are marine protected areas at enhancing commercial fish populations? Very often commissioners will be eager to ask for a list of possible interventions or actions, with the evidence of their effectiveness or understanding of the conditions under which one action is effective or not.
Appropriateness of a method	What is the most reliable method for monitoring changes in carbon stocks in forest ecosystems? Here the question aims to identify which of several methods would be the most appropriate to provide guidelines for users and policy.
Optimal management options	What is the optimal grazing regime for maximizing plant diversity in upland meadows? Such a concern relates to efficiency or cost-effectiveness of an intervention or combination ("bundle") of actions.
Optimal ecological or biological state	What is the desirable state of forest in terms of the distribution of deadwood and other biodiversity-relevant structures? This addresses values and philosophical approaches; the need for evidence would relate to the relationship between the state and its outcomes (e.g. ecosystem services)
Opinion or perception	Is there public support for badger culling in the UK? Datasets for this type of question may come from opinion polls or surveys, rather than experimental studies.
Ecological or geographical distribution	How has the distribution and abundance of rabies in fox populations changed in the last 10 years? Here one could ask if there is any evidence of change and whether it is homogeneous across spatial and temporal scales and species.

Box 2.1. Examples of initial concerns or problems that might generate potential questions for evidence synthesis but are not yet sufficiently well defined.

The impact of roads on wildlife
- Responses of invasive species to climate change and impact on native species
- Management of forest by local communities to achieve better biodiversity protection
- Influence of different levels of greenhouse gases on genetic mutation in terrestrial organisms
- Pollution of rivers and impact on the fecundity of fishes
- Resilience of different ecosystems to over-exploitation
- Effectiveness of reserves for preserving migrating species
- Improvement of agricultural practices to restore soil biodiversity
- Mitigation of effects of climate change by urban greening
- Efficiency of reintroductions, translocation and captive breeding programmes to restore populations
- Impact of educational programmes to protect endangered species
- How can we prevent contamination of native species by genetically modified organisms?

Initial questions arising from discussions of evidence needs are typically broad, sometimes complex, and possibly not well defined (termed 'open-framed' questions), whereas questions appropriate for Systematic Review are typically specific, well defined, and relatively simple (termed 'closed-framed' questions). Systematic Maps can often be better suited to broader questions. A discussion on how to progress from evidence need to evidence synthesis is provided in Section 2.3, with further explanation and examples of the difference between open-framed and closed-framed questions given in section 2.3.1.

2.2 Getting people involved

In progressing from evidence needs to consideration of a specific question and planning of a CEE Evidence Synthesis it is likely that several different groups will have an interest in being involved. The group of people that identify a need for evidence may not be the group that undertakes a synthesis (except where the question is entirely scientifically motivated). There are at least four definable, but not mutually exclusive, groups that could be involved in the conduct of a synthesis from this early stage:

The User Group (e.g. Client, Commissioner, Requester) – policy or practice groups that identify the need for evidence and might commission an Evidence Synthesis and/or use its findings in the context of their work.

The Review Team – the group that conducts the synthesis; the authors of the synthesis report. We retain the term 'Review Team' for convention but the terms 'Project Team' or 'Synthesis Team' could also be used.

The Stakeholder Group – all individuals and organisations that might have a stake in question formulation and findings of the synthesis.

CEE – the independent organisation that oversees the conduct, peer review and endorsement of the synthesis process and synthesis report based on these Guidelines and Standards.

Normally, to ensure independence of conduct and avoid conflicts of interest, any individual would not be a member of more than one of these groups (unless the potential conflict of interest has been accepted by all parties). Funding will often come from the User Group but can come from any one of these groups or be entirely independent. Funders must always be declared along with any other conflicts of interest that might arise (see Sections 4 and 10).

The User and the Stakeholder Groups will have a very important role in the formulation of the review question, and in its phrasing but should not be directly involved in conducting the Evidence Synthesis. They may also help to establish the list of sources of evidence and search terms (by providing some of them, or checking the list for completeness). Involving many people at an early stage may be particularly critical if the findings are likely to be contested (Fazey et al. 2004), such as in the site selection for establishment of windfarms. However, particularly for a Systematic Review, stakeholder input needs to be carefully managed to avoid the question becoming too broad, complex or just impossible to answer (Stewart & Liabo 2012). There are further opportunities throughout the review process for input from stakeholders, but as we shall see below, identifying and obtaining initial consensus on the question is crucial to the value of an evidence synthesis.

Systematic reviews and maps can greatly benefit from engaging with stakeholders to ensure that inputs and outputs are of the greatest relevance and reliability to all interested parties. Stakeholder engagement should reflect the broader methodology of systematic reviews, in that it should be conducted in a reliable, transparent way that aims to be as verifiable and objective as possible. Objectivity and repeatability can seem difficult goals when working with people who may have strong and diverse views. But by aiming for transparency and clarity, stakeholder engagement can be a reliable and verifiable process: these are key tenets of the parallel process of systematic review.

Various definitions of stakeholders exist in the literature, with perhaps the most widely cited one being “any group or individual who is affected by or can affect the achievement of an organisation’s objectives” (Freeman 1984). In their framework for stakeholder engagement, Haddaway et al. (2017) define systematic review stakeholders across three factors: actors, roles, and actions: in this way, one stakeholder group may have a diverse set of ways to engage with a review. Using a broad, encompassing definition of stakeholders can help to ensure that all relevant stakeholders are engaged, particularly minority groups.

The ways in which stakeholders can engage with a review are outlined in Figure 2.1. Stakeholder engagement can have a substantial impact on the reviewers, the review and the stakeholders, and reviewers should ensure they plan these activities carefully to ensure the do no harm and carefully consider the ethics of who and how they engage.

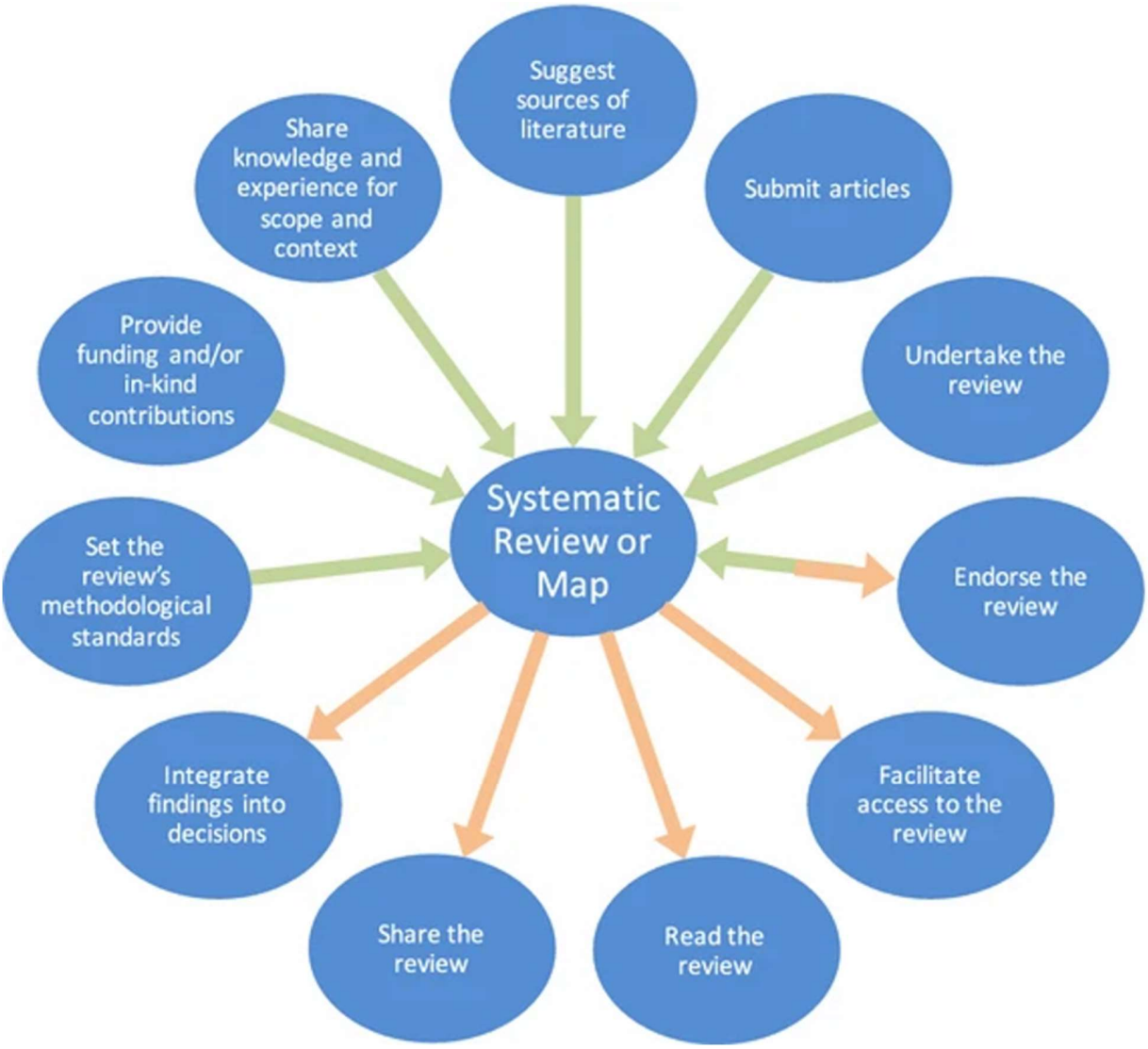


Figure 2.1. Model of potential benefits of stakeholder engagement. Models shows direction of benefit with respect to stakeholders (green arrows benefit the review, orange arrows benefit the stakeholders). https://environmentalevidencejournal.biomedcentral.com/articles/10.1186/s13750-017-0089-8/figures/2

Stakeholder engagement should be seen as a continual process that begins at the planning stages and continues through to communication of review/map findings. Careful and continued stakeholder engagement can help to increase the relevance and salience of environmental research and reviewers should carefully consider how to do so before submitting their protocols to CEE.

2.3 From a problem to a reviewable question: Question generation and formulation

As potential questions are generated and review teams are formed there will be a final process of question formulation. There is no set formal process for this but the critical elements are set out in this section.

Each Evidence Synthesis starts with a specific question whereas evidence needs are typically much broader. For commissioners and decision makers, finding the right question to inform decisions can be a compromise (probably more so in environmental sciences than in most other disciplines) between taking a holistic approach, involving a large number of variables and increasing the number of relevant studies, and a reductionist approach that limits the review's relevance, utility and value (Pullin et al. 2009). There can be a temptation to try to squeeze too much information out of one synthesis by including broad subject categories, multiple interventions or multiple outcome measures (this can be dealt with by first conducting a Systematic Map if time and resources allow). Equally, there may be a tendency to eliminate variables from the question so that a Systematic Review is feasible but its utility or 'real world' credibility (external validity - see Section 3.6.1) is limited. The formulation of the question is therefore of paramount importance for many reasons. For example:

- a Systematic Review question must be answerable using scientific methodology, otherwise relevant primary studies are unlikely to have been conducted
- the question should be generated by, or at least in collaboration with, relevant decision-makers (or organisations) for whom the question is real, to ensure its utility to inform.
- it may also be important for the question to be seen as neutral (unbiased) to stakeholder groups to minimise conflicts
- definitions of the structural elements of the question (see next section) are critical to the subsequent process because they generate the terms used in the literature search and determine relevance (eligibility) criteria.

The wording of the question and the definitions of question elements may be vital in establishing stakeholder consensus on the relevance of the synthesis. Ideally, meetings should be held with key stakeholders to try to reach consensus on the nature of the question. We recommend that experts in the field are consulted. Ideally, a meeting would invite some of them to present the state-of-the-art on the topic of interest so that each participant (especially Review Team members that are not subject experts) could be familiarised with the context, technical jargon and challenges.

2.3.1 Open-framed and closed-framed questions

As can be seen from the examples in Box 1, initial questions arising from discussions of evidence needs are often broad and may be rather "fuzzy" in that they often lack clear structural elements. For example, a question asking generally about the impact of roads on wildlife does not clarify what types of wildlife are of interest (e.g. terrestrial, aquatic, microorganisms, macro-organisms), their characteristics (e.g. individuals, populations, or communities), what types of impact are of concern (e.g. on abundance, dispersal, reproduction), or whether it matters what kinds of road (e.g. rural lanes, motorways) are considered. So, it would not be possible for a

Review Team to specify which types of evidence are needed to provide a meaningful answer to the question. Such questions may be suitable for Systematic Mapping but for Systematic Reviews to be feasible there needs to be sufficient structure in a question so that specific types of evidence relevant to the question (known as **key elements**) can be sought in order to answer it.

In the process of **question formulation** an initial, broad, policy question is broken down into more specific questions that are sufficiently well-structured to be amenable to Evidence Synthesis. In the case of the example of the impacts of roads on wildlife above, a refined version might be "what is the impact of motorways on populations of endemic bird species in Europe?" This question clearly has some structure to it now, as some key elements are specified, meaning that we can see which types of evidence relating to roads and birds we should be looking for. However, more thought needs to be given to this question in order for it to make full sense. If we are aiming to assess an impact of a motorway we would need to have a comparator condition as a reference. This key element is rarely explicitly specified in the question. But we could further refine the question, for example by stating that we would require primary studies that had compared bird-habitats containing motorways against similar bird-habitats without motorways. We would also need to specify the "outcome" key element (e.g. population abundance, or breeding success). The resulting question may then be: What is the impact of habitats containing motorways on the breeding success of endemic European birds, as compared to habitats without motorways? In principle, this question would be amenable to an Evidence Synthesis (Table 2.2).

Table 2.2 Example of question formulation: roads and wildlife

Question	Key elements	Question type
Starting question: What is the impact of roads on wildlife?	None specified other than roads (vague), wildlife (vague) and impact (vague)	Open-framed (possible for Systematic Mapping but unsuitable for Systematic Review)
Refined question: What is the impact of motorways on populations of endemic bird species in Europe?	Motorways (=exposure), European endemic bird species (=population); but comparator and outcome not specified	Open-framed (suitable for Systematic Mapping but unsuitable for Systematic Review)
Further refined question: What is the impact of habitats containing motorways on the breeding success of endemic European bird species, as compared to habitats without motorways?	Motorways (=exposure), no motorways (=comparator), European endemic bird species (=population), breeding success (=outcome)	Closed-framed (suitable for Systematic Mapping and possible for Systematic Review)

In the above example we started with a broad question with almost no structure and refined the question to a point where it contained structural elements, but not quite enough structure that

would permit a meaningful Evidence Synthesis. This type of question (i.e. lacking some or all of the required structural key elements) is known as an **open-framed question**. Such questions are normally not answerable in a single experimental study and therefore not answerable through an aggregative synthesis of similar studies. Further refinement of the open-framed question provides a well-structured question amenable for Systematic Review and, since all necessary key elements are now clearly specified, this is known as a **closed-framed question.**

Breaking down open-framed to identify closed-framed questions can be a valuable exercise in a policy context. The process can identify basic units of evidence that can be accumulated through Systematic Review and subsequently combined to inform higher-level decisions. Pullin et al. (2009) have outlined a process adapted from the health sciences. Essentially two stages are involved as outlined in Figure 2.1. The first requires that potential strategies for addressing open-framed questions are identified and the second that potential interventions are considered that would help deliver each strategy. The effectiveness of these interventions can then be the focus of a Systematic Review. Systematic Mapping can be used to inform the first stage and address an open-framed question, whereas Systematic Reviews might subsequently consider the effectiveness of individual interventions.

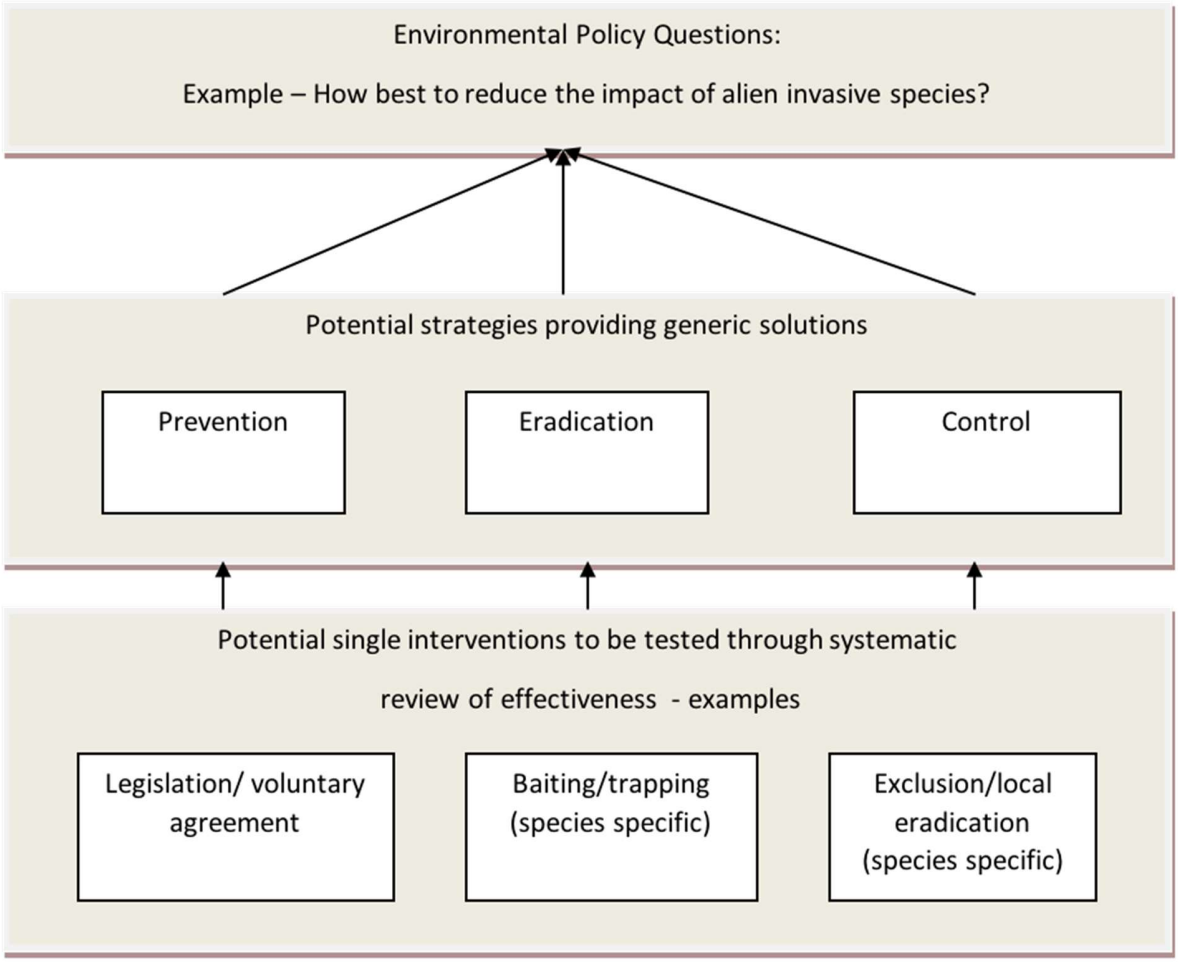


Figure 2.1. Relationship between a high-level open-framed policy question, potential generic solutions and individual interventions. After Pullin et al. (2009).

2.3.2 Key elements of questions amenable to Evidence Synthesis

The question illustrated in Table 2.2 is a comparative question. As such, it has four key elements which would need to be specified for it to be answerable (whether by a primary study or an Evidence Synthesis). These are: the population (P) of interest (endemic European birds); the exposure (E) of interest (motorways within habitat); the comparator (C) of interest (habitats without motorways); and the outcome (O) of interest (breeding success). This "PECO" type of question structure is very common. In cases where the exposure element is intentional, i.e. called an "intervention" (I), then the "PICO" acronym may be used instead, although PECO and PICO essentially indicate an identical question structure (Table 2.3).

Another example, illustrating a PICO-type question is '**are marine protected areas effective at conserving commercial fish stocks?'** In this case the key elements could be:

P = Populations of commercially important fish species,

I = Establishment of marine protected area,

C = Area with no protection or limited protection

O = Relative change in fish populations

Table 2.3. Elements of a reviewable PICO/PECO question: normally a permutation of 'does intervention/exposure I/E applied to populations of subjects P produce a measurable amount of outcome O when compared with comparator C?'.

Question element	Definition
Population (of subjects)	Unit of study (e.g. ecosystem, species) that should be defined in terms of the statistical populations of subject(s) to which the intervention will be applied.
Intervention/exposure	Proposed management regime, policy, action or the environmental variable to which the subject populations are exposed.
Comparator	Either a control with no intervention/exposure or an alternative intervention or a counterfactual scenario.
Outcome	All relevant outcomes from the proposed intervention or environmental exposure that can be reliably measured

Although Systematic Review methodology was initially developed to test the effectiveness of interventions in medical practice, its use has broadened considerably and the methodology is now also used to address a range of different types of questions that may have different key elements (Table 2.4). Other related question structures have been proposed and might be more applicable to some kinds of questions. SPICE (Setting, Perspective, Intervention, Comparator, Evaluation method) is an example that might be applicable to some questions suitable for CEE Systematic Reviews (see Booth 2004).

Table 2.4. Examples of questions amenable to Evidence Synthesis and broken down into their key elements.

Question Type	Question	Question Elements	Example Elements
		Population	Local human populations

Effect of intervention or exposure - "PECO" or "PICO" Structure	"What are the human wellbeing impacts of terrestrial protected areas?" (Pullin et al. 2012)	Intervention	Terrestrial protected areas/associated integrated development projects
		Comparator	Absence of PAs
		Outcome	Measures of human wellbeing
	What are the impacts of reindeer/caribou (Rangifer tarandus) on arctic and mountain vegetation?" (Bernes et al. 2013)	Population	Vegetation in alpine/subalpine areas and arctic/subarctic tundra
		Exposure	Herbivory by reindeer/caribou
		Comparator	No/less herbivory by reindeer/caribou
		Outcome	Vegetation change (assemblage or specific groups)
Analytical accuracy (diagnostic test accuracy) - "PIT" structure	"Comparison of methods for the measurement and assessment of carbon stocks and carbon stock changes in terrestrial carbon pools? (Petrokofsky et al. 2010)	Population	Forest ecosystems
		Test being evaluated (Index test)	Estimates of carbon content
		Target Condition	Carbon release or sequestration from ecosystem change
Prevalence, occurrence, incidence - "PO" structure	"What is the rate of occurrence of rabies in foxes in various European countries?"	Population	Red fox populations
		Outcome	Prevalence of rabies

Decision makers may often seek more than just an aggregate answer (e.g. a mean impact of an intervention) to the primary question. Secondary question elements, that follow on from the primary question, such as the cost-effectiveness of interventions; the prediction of reasons for variance in effectiveness (when or where will it work or not work?); the appropriateness and acceptability of particular interventions; and the factors which might influence the implementation of interventions 'in the real world' as opposed to the laboratory may be of equal or even greater importance. In many cases this might mean that the review essentially follows the 'effectiveness' review format but with development of synthesis strategies tailored to address a range of sub-questions. Discussion with funders and stakeholders is important at the beginning of

the process to identify the type of evidence needed, to assess whether or not an effectiveness type of Evidence Synthesis is the most appropriate.

Further examples of question formulation

Example 1:

Concern/Problem - Protected areas (PAs) must 'at least do no harm' to human inhabitants (Vth IUCN World Parks Congress, Durban 2003), but previously some PAs have been documented to have negative effects on humans living inside and around their borders. The Scientific and Technical Advisory Panel (STAP) of the Global Environment Facility (GEF) wanted to know how PAs affected human wellbeing and whether impacts had changed over time and with different governance structures.

Question Development - Terrestrial protected areas were considered distinct from marine in the context of human impacts. The Systematic Review would include established and new PAs and intrinsically linked development projects. All outcomes relating to measures of human wellbeing were deemed relevant. The commissioners decided that the target populations would include all local human populations living both within and around the PA, with 'local' being defined as broadly as up to and including a national level. A cutoff of 1992 was chosen for published studies, since all PAs had to conform to IUCN category guidelines established at the Convention on Biological Diversity in Rio de Janeiro, 1992.

Final Systematic Review Question - What are the human wellbeing impacts of terrestrial protected areas?

Lesson Learnt - With hindsight this question should be have been approached through systematic mapping, largely because of the many different ways that exist for measuring impact of human wellbeing.

Example 2:

Concern/Problem - Lowland peatland ecosystems constitute vast amounts of carbon storage relative to their geographical extent. Extraction and drainage of peat for fuel and agriculture can release greenhouse gases (GHG; CO_2, CH_4 and N_2O) and other carbon stores, contributing to global warming. Rewetting and wetland restoration aim to ameliorate these destructive practices but their effectiveness is uncertain. Whilst upland peat systems are relatively well-understood, no synthesis concerning lowland peats has been undertaken to date.

Question Development - The commissioners decided to focus the subject of a previous Systematic Review topic from all peatlands onto temperate and boreal regions, and widen the scope from water level changes to all changes in land management. Carbon fluxes and greenhouse gases were kept as relevant outcomes.

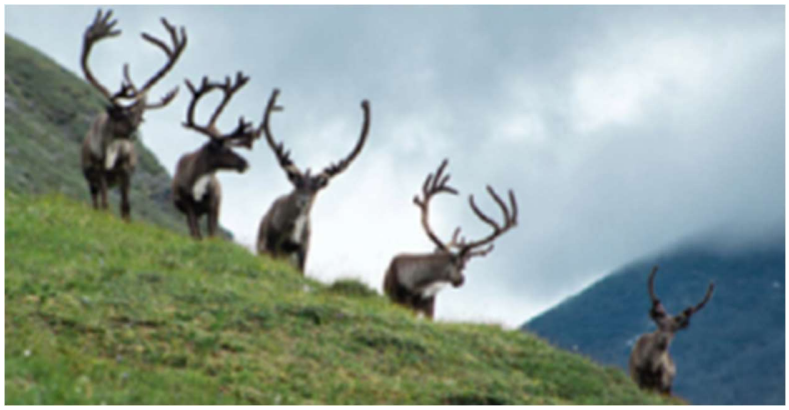
Final Systematic Review Question - How are carbon stores and greenhouse gas fluxes affected by different land management on temperate and boreal lowland peatland ecosystems?

Example 3:

Concern/Problem - What intensity of grazing should be recommended to conserve biodiversity whilst ensuring economic sustainability of reindeer herding? An early view that reindeer were responsible for overgrazing in northern parts of Scandinavia has changed, with current opinion being that the observed overgrazing was localized and short-lived. In contrast, some are now concerned that grazing levels are insufficient to control mountain vegetation. Stakeholders identified a need to clarify a vague political dogma and goal; that the Swedish mountains should be characterised by grazing.

Question Development - Development of the review question (initially suggested by the Swedish Environmental Protection Agency, SEPA) was undertaken by a team of scientists in consultation with stakeholders. Any impact resulting from herbivory by reindeer or caribou (both *Rangifer tarandus*) from anywhere in their natural or introduced range was chosen to be included in the scope of the review. Herbivory in coniferous forests was excluded, however, since the review was to be focused on mountain and arctic regions.

Final Systematic Review Question - What is the impact of reindeer/caribou (*Rangifer tarandus*) on mountain and arctic vegetation?



2.4 Systematic Review or Systematic Map?

The approaches to planning and conducting Systematic Reviews and Systematic Maps are similar in many ways but as forms of evidence synthesis they differ in their outputs. Systematic Reviews usually aim answer a question by collating and synthesising findings of individual studies in order to produce an aggregate measure of effect or impact. Systematic Maps do not aim to answer a specific question, but instead collate, describe and map findings in terms of distribution and abundance of evidence, often configured in relation to different elements of a question (Gough et al. 2012; James et al. 2016). As shown in Table 2.5, both methods share the same initial steps and differ primarily in their analytical approaches and outputs.

Table 2.5. Key aspects of Systematic Reviews and Systematic Maps

	Systematic Review	Systematic Map
Protocol	Mandatory	Mandatory
Systematic searching	Mandatory	Mandatory
Systematic study selection	Mandatory	Mandatory
Critical appraisal of study validity	Mandatory, to ensure robustness of the review answer – directly influences the data synthesis and interpretation steps	Optional (possible if study validity indicators can be captured using the coding method, but unlikely in practice) – does not influence mapping process itself
Data coding and extraction	Mandatory, Meta-data coded and outcome measures (e.g. effect sizes) extracted.	Mandatory, metadata only coded. No extraction of outcome measures (e.g. effect sizes).
Data synthesis approach	Aggregative, seeking an unbiased answer with known precision; could involve meta-analysis	Exploratory; may include coding and group analysis

Typical output	A quantitative or qualitative answer with an indication of uncertainty and any threats to validity. May include estimate of variance caused by external factors.	A description of the evidence base, showing the distribution and abundance of evidence across different elements of the question. A relational database may be provided.

When starting with a broad policy problem, it may not be immediately obvious whether a question can be developed that is suitable for Systematic Review or Systematic Mapping. This may become clearer as the question formulation process proceeds. If commissioners are seeking an answer that needs to be as precise as possible (or at least with uncertainty quantified) and free from bias this would signal the need for a Systematic Review. Alternatively, if commissioners are aware that the evidence base is broad and heterogeneous, they may seek information about the characteristics of the evidence base before deciding on how to proceed with their policy objectives. In such a situation it may be evident that the focus should be on conducting a Systematic Map rather than a Systematic Review.

The key characteristics of both CEE Systematic Reviews and Systematic Maps are rigour, objectivity and transparency. These characteristics serve to minimise bias and work toward consensus among stakeholders on the status of the evidence base.

There are different motivations for conducting Systematic Reviews and Systematic Maps. The latter are often preliminary syntheses of the evidence relating to a broader question that may subsequently inform more specific aggregative syntheses in the form of Systematic Reviews. Here we consider in more detail the decision makers' or commissioners' perspective and address the problem of deciding whether a Systematic Review or Systematic Map is the right option for informing their work.

Some examples of when a Systematic Review may be appropriate are when:

- There is a need to measure the effectiveness of an intervention or relative effectiveness of interventions.
- There is a need to measure the impact of an activity on a non-target population.
- There is a need to know the quantity and quality of research that has been conducted on a specific question.
- There are opposing views about the effectiveness of interventions or impact of actions.
- There is a need to consider the relative effectiveness and cost of interventions.

Systematic Maps may be a more suitable approach than Systematic Reviews when:

- A descriptive overview is required of the evidence base for a given topic
- The question is open-framed such as 'what interventions have been used to decrease the impact of commercial fishing on marine biodiversity?'

- The question is closed-framed but there are multiple subject populations, interventions and outcomes to consider such as 'what are the impacts of agri-environment schemes on farmland biodiversity?'
- Preliminary mapping is a useful stage to assess where evidence may be sufficient or lacking for further synthesis.

Systematic Reviews or Maps may not be appropriate when:

- The question is poorly defined or too complex (but see section 2.4)
- The question is too simple (e.g. has species x been recorded in region y)
- The question does not attract stakeholder (including scientific) interest (i.e. the rigour is not necessary)
- The question is not judged sufficiently important to be cost-effective to answer
- Very little good quality evidence exists but systematic confirmation of a knowledge gap will not be valued.
- A similar Systematic Review or Systematic Map has recently been completed (but see updating Section 1)
- The question can be satisfactorily answered with less rigorous and less costly forms of evidence synthesis

Ultimately the choice of Systematic Review or Systematic Map may come down to a matter of judgement of the Commissioners and/or Review Team. If Systematic Reviews are likely to include multiple syntheses, e.g. where multiple population types, interventions, or outcomes are defined in the question, an alternative option may be to conduct a Systematic Map first. This may identify subsets of the evidence base which are of highest priority (e.g. policy relevance) for subsequent focused synthesis by a Systematic Review. CEE will consider related Systematic Maps and Systematic Reviews on a case by case basis subject to the methods meeting adequate standards of rigour. Systematic Reviews which have multiple populations, interventions and/or outcome elements may contain a preparatory level of mapping as an integral element. However, this should not be an excuse for making the Systematic Review too speculative and the Protocol needs to be quite clear about how the review will proceed from the preparatory mapping to aggregative synthesis stages. Systematic Maps should not contain any aggregative synthesis elements. When a Systematic Map is followed by a separate Systematic Review, each requires its own separate Protocol.

In Systematic Mapping, the searching and inclusion processes are conducted with the same comprehensive method as for a Systematic Review, but the process does not extend to detailed critical appraisal or data synthesis. Data are, however, extracted from included studies in order to describe important aspects of the studies using a standard template and defined keywords and coding. This approach is designed to capture information on generic variables, such as the country in which a study took place, the population focus, study design and the intervention being assessed. This standard and well-defined set of keywords and codes is essential whenever classifying and characterising studies in order for reviewers to pull out key aspects of each study in a systematic way. For an example of a published CEE Systematic Map see Randall & James (2012). In this example, the Review Team examined the effectiveness of integrated farm

management, organic farming and agri-environment schemes for conserving biodiversity in temperate Europe. Their Systematic Map searched for relevant information in accordance with typical Systematic Review methodology. Screening of abstracts was then undertaken and a searchable database created using keywording to describe, categorise and code studies according to their focus and methodology. This searchable database is hosted on the CEE website and is freely available. Once the research has been mapped in this way it is then possible to identify pools of research which may be used to identify more narrowly defined review questions. For an example of this approach see Bowler et al. 2009.

2.5 Establishing a Review Team

Conducting a CEE Evidence Synthesis is a substantial piece of work and usually requires the input of a multidisciplinary team. Teams may consist of subject experts combined with review and synthesis methodology experts, such as information specialists or statisticians. Evidence Syntheses are normally undertaken by a team because one person is unlikely to possess all the skills required to conduct all stages of the review and synthesis, or have the appropriate combination of subject and methodological expertise, and because several stages of the review require independent conduct or checking that requires two or more participants to minimise the risk of introducing errors or bias. The Review Team should normally have a designated Lead Reviewer or Review Co-ordinator who is experienced in the methodology and a person (may also be the Lead Reviewer) able to project manage the rest of the team. The involvement of subject experts in the team brings with it the potential for bias. Careful consideration should be given to independence of subject experts within Review Teams and conflicts of interest should be declared, and avoided where possible.

It is preferable that the team is constituted before or during the establishment of the Protocol (Section 4) so that the team feels ownership and responsibility for its content. The rigorous methodology employed in CEE Evidence Syntheses requires substantial investment in time and it is important that careful planning of timetables and division of work is undertaken using some key predictors of the likely size and scope of the review. Loss of commitment among Review Team members when the workload becomes evident is a common cause of Evidence Syntheses stalling or failing to reach completion.

2.6 Involving stakeholders

Evidence Syntheses are driven by the question(s) they are trying to answer. Different people or organisations may view the question in different ways, perhaps from different ideological and theoretical perspectives. It is helpful therefore to involve a broad range of stakeholders at certain stages of an Evidence Synthesis so that different users' viewpoints are considered. These Guidelines recognise the many different opinions, academic and otherwise, about what constitutes a stakeholder (see also section 2.2 above). The term is used rather broadly here to mean people from organisations or representative groups who sit somewhere in the spectrum that indicates the extent to which they can affect (or be affected by) the issue being addressed by the Evidence Synthesis. One of the main aims of any Evidence Synthesis is to be transparent, and this includes being transparent about why a synthesis has the focus that it does. A broad stakeholder involvement will help to establish clearly that the Evidence Synthesis was carried

out in a way that attempted to remove the likely bias that would be introduced through a narrow, vested-interest, focus. Some of the types of stakeholders who have contributed to Evidence syntheses and should be considered when planning a synthesis are:

- Academics
- Government decision-makers (national, local)
- Intergovernmental decision makers
- Private sector – businesses, service providers
- Non-governmental or civil society organisations
- The general public

Stakeholders can have a very important role in framing the review question. They can also help to establish the list of sources of evidence and search terms (e.g. by providing some of them, or checking the list for completeness). Involving many people at an early stage may be particularly critical if the findings are likely to be contested (Fazey et al. 2004). Some Review Teams have found it useful to hold stakeholder workshops, usually at the question formulation stage, and sometimes also during the Protocol drafting stage of the synthesis, and the additional costs of such meetings should be considered during the planning stage.

2.7 Advisory Groups or Panels

Review Teams and Commissioners may wish to use an Advisory Group or Panel to help make decisions about the conduct of the review. Advisory groups can include methodological and subject area expertise, and include potential review users. They are Stakeholders who are willing to commit time to help the review group make necessary but difficult decisions in relation to the review topic at key times in its development to help ensure that the evidence synthesis is relevant and as free from bias as possible. They can be help particularly in framing the review question or refining the extent of a review once the size of the relevant literature becomes known. Such decisions can benefit from input from a variety of perspectives. However, Advisory Group members should be asked to declare conflicts of interest in the same way as Review team members. The role of any Advisory Group should be clearly specified in the Protocol, including at which points in the evidence synthesis they will be involved and how.

Section 3

Planning a CEE Evidence Synthesis

Last updated: March 15th 2021.

To meet CEE standards for the conduct of Evidence Syntheses the Review Team will need to establish an a priori Protocol detailing how they will conduct each stage of the Evidence Synthesis. The Protocol sets out how the question was formulated and how each stage of the synthesis will be conducted, and is submitted for approval and registration by CEE in advance of conducting the synthesis. The steps that aid planning the conduct of each stage are described in this section followed by guidance on the structure of the Protocol itself (Section 4). In addition,

a set of checklists (ROSES) that can be used during the preparation of any systematic evidence are available at https://www.roses-reporting.com/ . All the way through writing the protocol or final map/review, the checklists indicate the correct level of detail to be reported, so that the high standards of replicability are achieved.

3.1 Scoping the evidence

Before the commencement of an Evidence Synthesis, it is essential that some 'scoping' is undertaken to guide the construction of a comprehensive and appropriate Protocol, and to provide an indication of the likely form of the synthesis and thus facilitate resource planning. In certain circumstances, it may not be efficient to commit to a synthesis without some prior estimation of its value in terms of the likely extent and reliability of its findings. In addition, when scoping a Systematic Review, an estimate of the type of data (quantitative, qualitative) may be desirable to inform the type of data synthesis that might be appropriate.

Scoping may be undertaken by the commissioning organisation, by the Review Team, or a combination of the two. A thorough scoping exercise might entail:

- The development and testing of a search strategy (see below).
- An estimate of the volume of relevant literature and the volume of material likely to be unavailable in easily-accessible format (see below).
- An estimate of resources required based on the above, including time and personnel to achieve the search and sorting of the literature, possible financial resources to obtain some articles, contact some authors, use the help of translators, and even plan for possible need of statisticians if quantitative data are identified during this scoping stage.
- An estimate of the study types likely to be found (as identified through focused data extraction of a small subset of relevant papers). This may indicate whether a meta-analysis will be possible (for Systematic Review only).

The expected output from a scoping exercise is an estimate of the quantity of evidence, and a characterisation of the likely evidence base, pertaining to the question (see Box 3.1 for an example). The extent of investment in scoping required to meet CEE standards will differ with each Evidence Synthesis. We detail below the steps of a full scoping exercise.

3.2 Developing and testing a search strategy

Systematic and comprehensive searching for relevant studies is essential to minimise bias (see Section 5). The searching step requires more planning and preparation than other stages and so most of this Section is devoted to this task. Enlisting an information specialist in the review team is recommended so that an efficient search strategy can be established. Aside from validity, a good search strategy can make a substantial difference to the time and cost of a synthesis. A step-by-step overview of the search process for evidence synthesis is illustrated in Figure 3.1

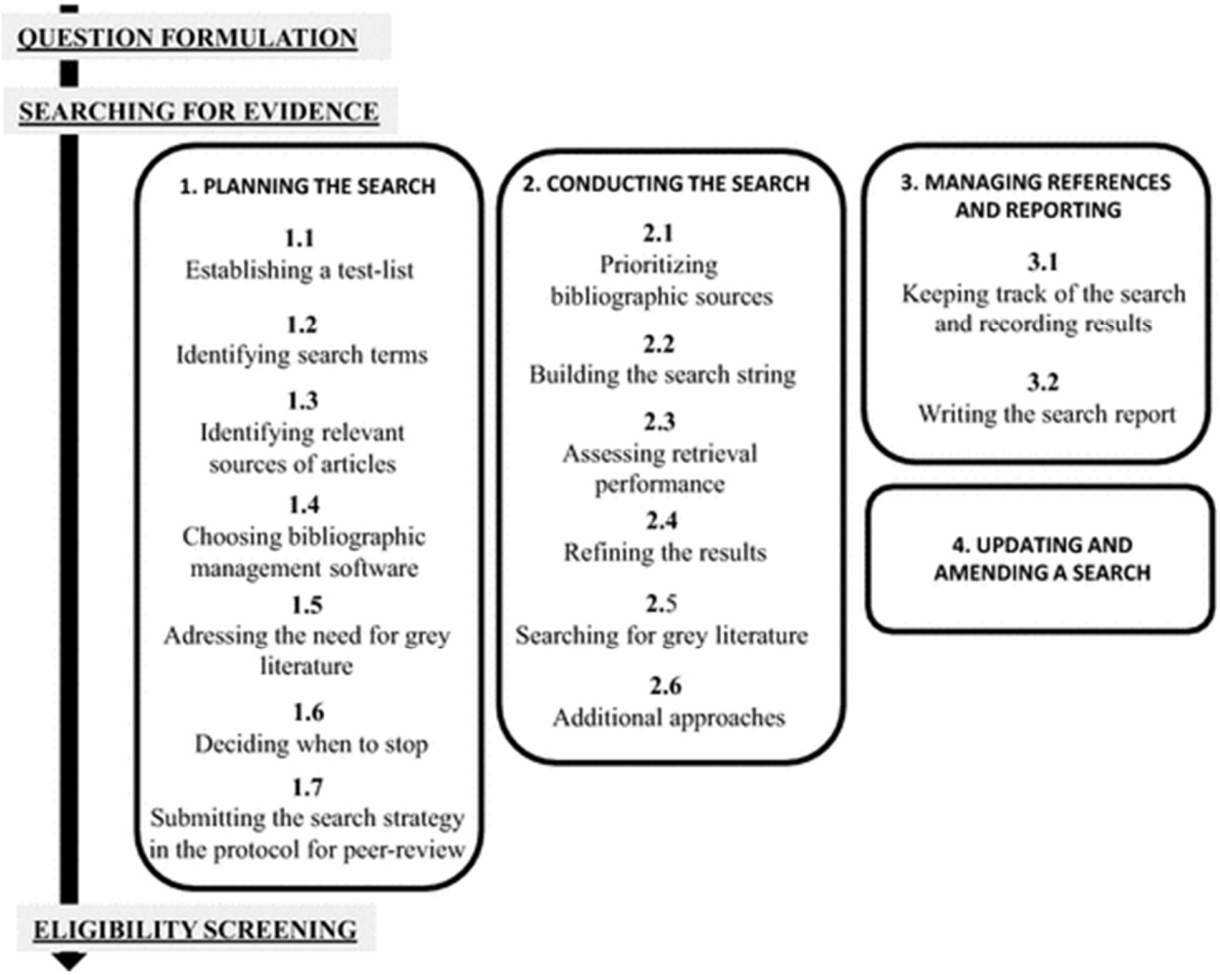


Figure 3.1 A guide to the planning, conduct, management and reporting of the searching phase of systematic reviews and systematic maps (after Livoreil et al. 2017).

In practice, it is unlikely that absolutely all of the relevant literature can be identified during an evidence synthesis search, for several reasons: (1) literature is often searched and examined only in those languages known to the project team; (2) some articles may not be accessible due to restricted access pay walls or confidentiality; (3) others lack an abstract or have unhelpful titles, which makes them difficult to identify; (4) others may simply not be indexed in a searchable database. Within these constraints, searches conducted for evidence synthesis should be as comprehensive as possible, and they should be documented so they can be repeated and readers can appreciate their strengths and weaknesses. Reporting any limitations to searches, such as unavoidable gaps in coverage (e.g. lack of access to some literature) is an important part of the search process, to ensure that readers have confidence in the review methods, and to qualify the interpretation of the evidence synthesis findings.

Steps involved in planning a search are presented in chronological order, bearing in mind that some of the process may be iterative. We also highlight the methods that enable the project team to identify, minimise and report any risks of bias that may affect the search and how this can affect the findings of an evidence synthesis.

We *use the following terminology:* **search terms** encompasses individual or compound words used in a search to find relevant articles. A **search string** is a combination of search terms combined using Boolean operators. A **search strategy**is the whole search methodology, including search terms, search strings, the bibliographic sources searched, and enough information to ensure the reproducibility of the search. **Bibliographic sources** (see below for more details) capture any source of references, including electronic bibliographic databases, those sources which would not be classified as databases (e.g. the Internet via search engines), hand searched journals, and personal contacts.

Preventing errors and biases

Conducting a rigorous evidence synthesis implies to try to minimise risks of errors and biases which may happen at all stages. Errors that can occur during the search include: missing search terms, unintentional misspelling of search terms, errors in the search syntax (e.g. inappropriate use of Boolean operators, see below) and inappropriate search terms. Such problems may be minimised when the search term identification process is conducted rigorously, and by peer-reviewing the search strategy, including within and outside the project team, during development of the Protocol (See Section 4).

Biases (systematic errors) in the search strategy may affect the search outcomes (Song et al. 2010). The methods used to minimize bias should be reported in the Protocol and the final review or map. Minimizing bias may require 1) looking for evidence outside traditional academic electronic bibliographic sources (e.g. grey literature); 2) using multiple databases and search tools to reduce the possibility of bias in the retrieved results; and, 3) contacting organisations or individuals who may have relevant material (Bayliss & Beyer 2015). Some biases have been listed in Bayliss & Beyer (2015) and a few of them are reported here to be considered by project teams as appropriate: **language bias** (Song et al., 2010) means that studies with significant or 'interesting' results are more likely to be published in the English language and easier to access to than results published in other languages. The impacts of this on synthesis outcomes have been evaluated, and consequences of omitting non-English-language studies could be serious (e.g. providing a different direction of mean effect; Konno et al. 2020). The way to reduce the risk of language bias is to look beyond the English language literature. **Prevailing paradigm bias** (Bayliss & Beyer, 2015) suggests that studies relating to or supporting the prevailing paradigm or topic (for example climate change) are more likely to be published and hence discoverable. To reduce this bias Review Teams should not rely only on finding well known relevant studies. **Temporal bias** includes the risk that studies supporting a hypothesis are more likely to be published first (Bayliss & Beyer, 2015). The results may not be supported by later studies (Leimu and Koricheva, 2004). Due to the culture of 'the latest is best', older articles may be overlooked and mis-interpretations perpetuated. The ways to reduce this bias include searching older publications, considering updating the search in the future, or test statistically whether this bias significantly affects the results of studies. **Publication bias** (Dickersin, 2005; Hopewell et al. 2007; Song et al., 2010) refers to asymmetry in the likelihood of publishing results: statistically significant results (positive results) are more likely to be accepted for publication than non-significant ones (negative results). This has been a source of major concern for systematic reviews and meta-analysis as it might lead to overestimating an effect/impact of an Intervention or Exposure on a Population (e.g. Gurevitch & Hedges, 1999; Rothstein et al.

2005; Lortie et al. 2007). To minimise this bias, searches for studies reporting non-significant results (most probably found in grey literature and studies in languages other than English) should be conducted in all systematic reviews and maps (Leimu & Koricheva 2005).

Structuring the search with PICO/PECO elements

An evidence synthesis process starts with a question that is usually structured into "building blocks" (concepts or elements), some of which are then used to develop the search strategy. The search strategy illustrated below is based on PICO/PECO elements which are commonly used in CEE evidence synthesis. Other elements and question structures exist (See Section 2). In any of these question structures it is possible to narrow the question (and the search) by adding additional search terms defining the Context or Setting of the question (e.g. "tropical", "experimental", or "pleistocene"). Searching for geographic location is not recommended because location names may be difficult to list or duplicate when the geographical range is broad. Geographical elements (e.g. name of the country) may, instead, be more efficiently used as eligibility screening criteria (see below).

Use of multiple languages

Identifying which languages are most relevant for the search may depend on the topic of the evidence synthesis. There are two main challenges with languages for an evidence synthesis: translating search terms into various languages to capture as many relevant articles as possible, and then being able to select and use the paper when not written in a language spoken by the project team members. In many electronic bibliographic sources, articles written in languages other than English can be discovered using English search terms. However, a large literature in languages other than English remains to be discovered in national and regional databases, e.g. CiNii for Japanese research. Searching is likely to require a range of languages when relevant articles are produced at national level, as much of it will be published in the official language of those nations (Corlett, 2011). Reporting the choice of language(s) in the Protocol and in the final synthesis report is important to enable repetition and updating when appropriate.

Human resources needed for searching

Each evidence synthesis is conducted by a project team. It may be composed of a project leader and associated experts (thematic and methodological). Because of the systematic aspect of the searching and the need to keep careful track of the findings, project teams should, when possible, include librarians or information specialists. Subject specialist librarians are conversant with bibliographic sources, and are often very familiar with the nuances of different transdisciplinary and subject-specific resources (Zhang et al. 2006). They are aware of the broad range of tools available for undertaking literature searches and they are aware of recent improvements in the range and use of those tools. They are also expert in converting research questions into search strategies. Such experts can themselves benefit by contributing to a project team since their institutions may require demonstration of collaborative work (Holst et al. 2005).

3.2.1 Planning the search strategy

The first step in planning a search is to design a strategy to maximise the probability of identifying relevant articles whilst minimizing the time spent doing so. Planning may also include discussions about eligibility criteria for subsequent screening (Frampton et al. 2017) as they are often linked to search terms. Planning should also include discussions about decision criteria defining when to stop the search as resource constraints (such as time, manpower, skills) may be a major reason to limit the search and should be anticipated and explained in the Protocol.

Establishing a test-list

A **test-list** is a set of articles that have been identified as relevant to answer the question of the evidence synthesis (e.g. are within the scope and provide some evidence to answer the question). The test-list can be created by asking experts, researchers and stakeholders (i.e. anyone who has an interest in the review question) for suggestions and by perusing existing reviews. The project team should read the articles of the test-list to make sure they are relevant to the synthesis question. Establishing a test-list is independent of the search itself and is used to help develop the search strategy and to assess the performance of the search strategy. The performance of a search strategy should be reported, i.e. whether the search strategy correctly retrieves relevant articles and whether all available relevant literature to answer the evidence synthesis question is likely to have been identified. The test-list may be presented in the Protocol submitted for peer-review.

The test-list should ideally cover the range of authors, journals, and research projects within the scope of the question. In order to be an effective tool it needs to reflect the range of the evidence likely to be encountered in the review. The number of articles to include in the test-list is a case-by-case decision and may also depend on the breadth of the question. When using a very small test-list, the project team may inappropriately conclude that the search is effective whilst it is not. Using the test-list may be an indicator for the project team to improve the search strategy, or to help decide when to stop the search.

Identifying search terms

A search string that is efficient at finding relevant articles means that a maximum of relevant papers will have been found and the project team will not have to run the search again during the course of the conduct of the evidence synthesis. Moreover, it may be re-used as such when amending or updating the search in the future, saving time and resources. Initial search terms can usually be generated from the question elements and by looking at the articles in the test-list. However, authors of articles may not always describe the full range of the PICO/PECO criteria in the few words available in the title and abstract. As a consequence, building search strings from search terms requires project teams to draw upon both their scientific expertise, a certain degree of imagination, and an analysis of titles and abstracts to consider how authors might use different terminologies to describe their research.

Reading the articles of the test-list as well as existing relevant reviews often helps to identify search terms describing the population, intervention/exposure, outcome(s), and the context of

interest. Synonyms can also be looked for in dictionaries. An advantage of involving librarians in the project team and among the peer-reviewers is that they bring their knowledge of specialist thesauri to the creation of search term lists. For example, for questions in agriculture, CAB Abstracts provides a thesaurus whose terms are added to database records. The thesaurus terms can offer broad or narrow concepts for the search term of interest, and can provide additional ways to capture articles or to discover overlooked words (http://www.cabi.org/cabthesaurus/). As well as database thesauri that offer terms that can be used within individual databases, there are other thesauri that are independent of databases. For example, the Terminological Resource for Plant Functional Diversity (http://top-thesaurus.org/) offers terms for 700 plant characteristics, plant traits and environmental associations. Experts and stakeholders may suggest additional keywords, for instance when an intervention is related to a special device (e.g. technical name of an engine, chemical names of pollutants) or a population is very specific (e.g. taxonomic names which have been changed over time, technical terminology of genetically-modified organisms). Other approaches can be used to identify search terms and facilitate eligibility screening (e.g. text-mining, citation screening, cluster analysis and semantic analysis) and are likely to be helpful for CEE evidence synthesis.

The search terms identified using these various methods should be presented as part of the draft evidence synthesis Protocol so that additional terms may be suggested by peer-reviewers. Once the list is finalised in the published Protocol it should not be changed, unless justification is provided in the final evidence synthesis report.

Developing search strings

The development of effective search strings (combinations of key words and phrases) for searching should take place largely during the planning stage, and will most likely be an iterative process, testing search strings using selected databases, recording numbers of references identified and sampling titles for proportional relevance or **specificity** (the proportion of the sample that appears to be relevant to the Evidence Synthesis question). **Sensitivity** (the proportion of potentially relevant articles identified as estimated using the test list) should improve as testing progresses and reach 100% when results from databases are combined. The iterative process may include considering synonyms, alternative spellings, and non-English language terms within the search strategy. An initial list of search terms may be compiled with the help of the commissioning organisation and stakeholders. All iterations of tested terms should be recorded, along with the number of references (hits) they return. This should be accompanied by an assessment of proportional relevance, so that the usefulness of individual search terms can easily be examined. Comparing search results when you include or exclude particular terms will allow you to identify superfluous or ineffective terms, and work out whether any should be removed from your search strategy. It is important to remember, however, that the functionality of different literature databases may vary considerably and terms that are apparently useful in one source will not always be appropriate in others: thus, search strings may need to be modified to suit each one.

Boolean operators (AND, OR, NOT) specify logic functions. They are used to group search terms into blocks according to the PICO or PECO elements, so that the search is structured and easy to understand, review and amend, if necessary. AND and OR are at the core of the structure

of the search string. Using AND decreases the number of articles retrieved whilst using OR enlarges it, so combining these two operators will change the exhaustivity and precision of the search.

OR is used to identify bibliographic articles in which at least one of the search terms is present. OR is used to combine terms within one of the PICO elements, for example all search terms related to the Population. Using "forest OR woodland OR mangrove" will identify documents mentioning at least one of the three search terms.

AND is used to narrow the search as it requires articles to include at least one search term from the lists given on each side of the AND operator. Using AND identifies articles which contain, for example, both a Population AND an Intervention (or Exposure) search term. For instance, a search about a population of butterflies exposed to various toxic compounds and then observed for the outcomes of interest can be structured as three sets of search terms combined with AND as follows: "(lepidoptera OR butterfly OR coleoptera OR beetle) AND (toxi* OR cry* OR vip3* OR Bacillus OR bt) AND (suscept* OR resist*)". Note: truncating words with* (see Box 3.1) at 3 characters (e.g. cry* in this example) may find lots of irrelevant words and may not be recommended.

NOT is used to exclude specified search terms or PICO elements from search results. However, it can have unanticipated results and may exclude relevant records. For this reason, it should not usually be used in search strategies for evidence synthesis. For example, searching for 'rural NOT urban' will remove records with the word 'urban', but will also remove records which mention both 'rural' AND 'urban'.

Proximity operators (e.g. SAME, NEAR, ADJ, depending on the source) can be used to constrain the search by defining the number of words between the appearance of two search terms. For example, in the Ovid interface "pollinators adj4 decline*" will find records where the two search terms "pollinators" and "decline" are within four words of each other. Proximity operators are more precise than using AND, so may be helpful when a large volume of search results are being returned.

Box 3.1. Example of test search

A test search was undertaken as part of the development of a Systematic Review protocol for the question "What is the evidence that scarcity and shocks in freshwater resources cause conflict instead of promoting collaboration?" (Johnson et al. 2011). This process involved the trialling and refining of search terms in the scientific citation indexing service 'Web of Science' based on a full list of relevant exposure and outcome terms identified through an initial scoping and discussion with the expert Review Team. The initial sets of search terms were as follows (asterisks (*) indicate that a truncation character was used to ensure that different spellings or synonyms would be captured – e.g. water* would also capture "waters", "watercourse(s)" and "waterway(s)" plus any other terms with the "water" word stem):

Pre-scoping exposure terms	water*, riparian*, aquifer*, aqua*, dam, dams, hydrolog*, hydroelectric*, groundwater, drought*, river*, lake*, stream, streams, reservoir*, flood*, irrigat*, rain*, baseflow*, precipitation, fresh*, basin*, flow, drylands
Pre-scoping outcome terms	conflict*, dispute*, insurgen*, war*, violen*, securit*, terror*, strife, peace*, govern*, coercion, cooperat*, "co-operat*", collaborat*, collective, geopolitic*, "international relation*" allocat*, distribut*, shar*, mediat* governance, treaty, treaties, agreement*, manag*

When testing the search terms, each of the 24 exposure terms above was tested individually with the outcome terms conflict* OR cooperat* and the first 100 articles returned were screened to assess the exposure search term's usefulness (after sorting by relevance). Once the 24 terms had been refined and finalised, these terms were searched with each outcome search term individually in a similar way to produce a final search string.

Search Terms	References	Comments
[all exposure terms separated by 'OR'] AND (conflict* OR cooperat*)	>100,000	Large number of unrelated articles – i.e. low specificity
(rain OR rains OR rainfall) AND (conflict* OR cooperat*)	604	Rain* changed to (rain OR rains OR rainfall), and resulted in relevant hits. Retained in final search string.
baseflow* AND (conflict* OR cooperat*)	3	No relevant articles; 'baseflow' was therefore excluded from the final search.

Test searches were undertaken using three bibliographic services and databases: ISI Web of Knowledge (now called Web of Science), OCLC First Search (now integrated into WorldCat Discovery services) and Science Direct. In addition, a web-based search for grey literature was trialled using the search engine Google. Furthermore, professional organisations were identified and additional articles were provided by the Review Team and stakeholders with a personal knowledge of the topic.

All test searches should be carefully recorded (including the date of the search) and saved so that they may be accessed later, removing duplication of effort where possible. However, since the test searches are conducted in advance of the actual search, it will be necessary to update the search again in order to check whether any recent literature has become available. In the larger bibliographic databases and services it is possible to save searches and set up an alert service that will periodically run the saved searches and return new records. This can be useful if the testing occurs well in advance of the synthesis, or if the synthesis runs over a long period of time.

A high-sensitivity and low-specificity approach is often necessary to capture all or most of the relevant articles available, and reduce bias and increase repeatability in capture (see below). Typically, large numbers of articles are therefore identified but rejected at the title and/or abstract screening stage.

A final step in the development of the search terms and strings is to test the strategy with the test list. A comprehensive set of terms and strings with an appropriate balance of specificity and sensitivity SHOULD retrieve these relevant articles without returning an unmanageable number of irrelevant articles. Reasons why any articles from the test list were not retrieved should be investigated so that the search strategy can be appropriately modified to capture them.

The Review Team should report the performance of the search strategy in the Protocol with an update in the final report (e.g. as a percentage of the test-list finally retrieved by the search strategy when applied in each electronic bibliographic source, e.g. Söderström et al. 2014, Haddaway et al. 2015). A high percentage is one indicator that the search has been optimized and the conclusions of the review rely on a range of available relevant articles that reflect at least those provided by the test-list. A low percentage would indicate that the conclusion of the review could be susceptible to change if other 'missed' articles are added. The test list should be fully captured when searches from all bibliographic sources are combined.

Assessing the volume of literature

The volume of literature arising from test searches may be used as a predictor of the extent of the evidence base and a crude predictor of its strength (number of rigorous studies). For example, whether the review question is likely to identify a knowledge gap (very few articles), seems too broad and should be broken down or targeted toward a systematic map approach (very many diverse articles covering a range of populations, interventions and/or outcomes), or if it has the potential to provide an answer to the question as it is currently phrased and with the resources highlighted by the scoping exercise (nothing needs to be changed). This has implications in terms of the time and resources required to complete the review. Note, however, that the total number of returned articles is likely to reflect the specificity of the chosen search terms (and possibly searching skills of the Review Team) and is only an indicator. This can then be used to extrapolate and determine the likely quantity (but not quality) of articles relevant to the review question. The volume of literature that is likely to be difficult to access (in languages unfamiliar to the Review Team, or in publications that are not available electronically or not readily available in libraries) should, if possible, be assessed at this stage.

Identifying relevant sources of articles

Various sources of articles relevant to the question may exist. Understanding the coverage, the functions and limitations of information sources can be time-consuming, so involving a librarian or information specialist at this stage is highly recommended. We will use **bibliography** to refer to a list of articles generally described by authorship, title, year of publication, place of publication, editor, and often, keywords as well as, more recently, DOI identifiers.
A **bibliographic source** allows these bibliographies to be created by providing a search and retrieval interface. Much of the information today is likely to come from searches of **electronic bibliographic sources**, which are becoming increasingly comprehensive with the passage of time as more material is digitised. Here we use the term **electronic bibliographic source** in the broad sense. It includes individual electronic bibliographic sources (e.g. Biological Abstracts) as well as platforms that allow simultaneous searches of several sources of information (e.g. Web of Science) or could be accessed through search engines (such as Google). Platforms are a way to access databases.

Coverage and accessibility

Several sources should be searched to ensure that as many relevant articles as possible are identified (Avenell et al., 2001; Grindlay et al. 2012). A decision needs to be made as to which sources would be the most appropriate for the question. This mostly depends on the disciplines addressed by the question (e.g. biology, social sciences, other disciplines) and the identification of sources that may provide the greatest quantity of relevant articles for a limited number of searches and their contribution in reducing the various biases described earlier in the paper (see 1.3). The quantity of results given by an electronic bibliographic source is NOT a good indicator of the relevance of the articles identified and thus should not be a criterion to select or discard a source. Information about access to databases and articles (coverage) can be obtained directly from the project team by sharing knowledge and experience, asking librarians and information experts and, if needed, stakeholders. Peer-review of the evidence synthesis Protocol may also provide extra feedback and information regarding the relevance of searching in some other sources.

Some sources are open-access, such as Google Scholar, whereas others require subscription such as Scopus. Therefore, access to electronic bibliographic sources may depend on institutional library subscriptions, and so availability to project teams will vary across organisations. A diverse project team from a range of institutions may therefore be beneficial to ensure adequate breadth of search strategies. When the project team does not have access to all the relevant bibliographic sources, it should explain its approach and list the sources that were available but not searchable and acknowledge these limitations. This may include indications as to how to further upgrade the evidence synthesis at a later stage.

Types of sources

In this subsection we first present bibliographic sources which allow the use of search strings, mostly illustrated from the environmental sciences. An extensive list of searchable databases for the social sciences is available in Kugley et al. (2016). Other sources and methods mentioned

below (such as searches on Google) are complementary but cannot be the core strategy of the search process of an evidence-synthesis as they are less reproducible and transparent.

Bibliographic sources may vary in the search tools provided by their platforms. Help pages give information on search capabilities and these should be read carefully. Involving librarians who keep up-to-date with developments in information sources and platforms is likely to save considerable time.

Electronic bibliographic sources

The platforms which provide access to bibliographic information sources may vary according to:

A) Platform issues

- the syntax needed within search strings (see 2.2) and the complexity of search strings that they will accept
- access: not all bibliographic sources are completely accessible. It depends on the subscriptions available to the project team members in their institutions. The Web of Science platform, for example, contains several databases, and it is important to check and document which ones are accessible to the project team via that platform.

B) Database issues

- disciplines: subject-based bibliographic sources (CAB ebooks; applied life sciences, agriculture, environment, veterinary sciences, applied economics, food science and nutrition) versus multidisciplinary sources (Scopus, Web of Science);
- geographical regions (e.g. Latin America, HAPI-Hispanic American Periodicals Index, or Europe CORDIS). It may be necessary to search region-specific bibliographic sources if the evidence-synthesis question has a regional focus (Bayliss & Beyer, 2015);
- document types: scientific papers, conference or proceedings, chapters, books, theses. Many university libraries hold digital copies of their theses, such as the EThOS British Library thesis database. Conference papers may be a source of unpublished results relevant for the synthesis, and may be found through the BIOSIS Citation index or the Conference Proceedings Citation Index (Thomson Reuters 2016, in Glanville, in press)
- durations at the time of writing, in the Web of Science Core Collection some articles may be accessible from 1900 although by no means all, in Scopus they may date from 1960);

Publishers' databases

The websites of individual commercial publishers may be valuable sources of evidence, since they can also offer access to books, chapters of books, and other material (e.g. datasets). Using their respective search tools and related help pages allows the retrieval of relevant articles based on search terms. For example, Elsevier's ScienceDirect and Wiley Interscience are publishers'

platforms that give access to their journals, their tables of contents and (depending on licence) abstracts and the ability to download the article.

Web-based search engines

Google is one example of a web-based search engine that searches the Internet for content including articles, books, theses, reports and grey literature (see 1.5 and 2.5 Grey literature). It also provides its own search tools and help pages. Such resources are typically not transparent (i.e. they order results using an unknown and often changing algorithm, Giustini & Boulos, 2013) and are restricted in their scope or in the number of results that can be viewed by the user (Google Scholar). Google Scholar has been shown not to be suitable as a standalone resource in systematic reviews but it remains a valuable tool for supplementing bibliographic searches (Bramer et al. 2013; Haddaway et al. 2015) and to obtain full-text PDF of articles. BASE Bielefeld academic search engine (https://www.base-search.net) is developed by the University of Bielefeld (Germany) and gives access to a wide range of information, including academic articles, audio files, maps, theses, newspaper articles, and datasets. It lists sources of data and displays detailed search results so that transparent reporting is facilitated (Ortega 2004).

Full-text documents will be needed only when the findings of the search have been screened for eligibility and retained based on their title and abstract, and need to be screened at full-text (see Frampton et al. 2017). Limited access to full-texts might be a source of bias in the synthesis, and finding documents may be time-consuming as it may involve inter-library loans or direct contact with authors. Documents can be obtained directly if (a) the articles are open-access, (b) the articles have been placed on an author's personal webpage, or (c) are included in the project team' institutional subscriptions. Checking institutional access when listing the sources of bibliography may help the project team anticipate needs to get extra support.

Choosing bibliographic management software

Bibliographic searches may produce thousands or sometimes tens of thousands of references that require screening for eligibility and so it is important to ensure that search results are organised in such a way that they can be screened efficiently for their eligibility for an evidence synthesis. Key actions that will be necessary before screening can commence are to assemble the references into a library, using one or more bibliographic reference management tool(s); and to identify and remove any duplicate references.

Assembling references

A range of bibliographic reference management tools are available into which search results may be downloaded directly from bibliographic databases or imported manually, and these vary in their complexity and functionality. Some tools, such as Eppi Reviewer (Social Science Research Unit, 2016) and Abstrackr (Rathbone et al. 2015) include text mining and machine learning functionality to assist with some aspects of eligibility screening. According to recently-published evidence syntheses and Protocols, the most frequently-used reference management tools in CEE evidence syntheses are Endnote and Eppi Reviewer (sometimes used in combination with Microsoft Excel), although others such as Mendeley and Abstrackr are also used. Given that

reference management tools have diverse functionality and are continually being developed and upgraded, it is not possible to recommend any one tool as being 'better' than the others. An efficient reference management tool should:

- enable easy removal of duplicate articles, which can reduce substantially the number of articles;
- readily locate and import abstracts and full-text versions for articles where available;
- enable the review team to record their screening decisions for each article;
- enable articles, and any screening decisions accompanying them, to be grouped and analysed to assist the team in checking progress with eligibility screening and in identifying any disagreements between screeners.

Other features of reference management tools that review teams may find helpful to consider are: whether the software is openly accessible (e.g. Mendeley) or may require payment (e.g. Endnote, Eppi Reviewer); the number of references that can be accommodated; the number of screeners who can use the software simultaneously; and how well suited the tool is for project management tasks, such as allocating eligibility screening tasks among the review team members and monitoring project progress.

Addressing the need for grey literature

"**Grey literature**" relates to documents that may be difficult to locate because they are not indexed in usual bibliographic sources (Konno & Pullin 2020). It has been defined as *"manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by libraries and institutional repositories, but not controlled by commercial publishers; i.e. where publishing is not the primary activity of the producing body" (12th Int Conf On Grey Lit. Prague 2010, but see Mahood et al., 2014).* Grey literature includes reports, proceedings, theses and dissertations, newsletters, technical notes, white papers, etc. (see list on http://www.greynet.org/greysourceindex/documenttypes.html). This literature may not be as easily found by internet and bibliographic searches, and may need to be identified by other means (e.g. asking experts) which may be time-consuming and requires careful planning (Saleh et al. 2014).

Searches for grey literature should normally be included in evidence synthesis for two main reasons: 1) to try to minimize possible publication bias (Hopewell et al. 2007), where 'positive' (i.e. confirmative, statistically significant) results are more likely to be published in academic journals (Leimu and Koricheva 2005); and 2) to include studies not intended for the academic domain, such as practitioner reports and consultancy documents which may nevertheless contain relevant information such as details on study methods or results not reported in journal articles often limited by word length.

Deciding when to stop

If time and resources were unlimited, the project team should be able to identify all published articles relevant to the evidence-synthesis question. In the real world this is rarely

possible. Deciding when to stop a search should be based on explicit criteria and it should be explained in the Protocol and/or synthesis report. Often, reaching the budget limit (in terms of project team time) is the key reason for stopping the search (Saleh et al. 2014) but justification for stopping should rely primarily on the acceptability of the performance of the search for the project team. Searching only one database is not considered as adequate (Kugley et al. 2016). Observing a high rate of article retrieval for the test-list should not preclude the conduct additional searches in other sources to check whether new relevant papers are identified. Practically, when searching in electronic bibliographic sources, search terms and search strings are modified progressively, based on what is retrieved at each iteration, using the "test-list" as one indicator of performance. When each additional unit of time spent in searching returns fewer relevant references, this may be a good indication that it is time to stop the search (Booth 2010). Statistical techniques, such as capture-recapture and the relative recall method, exist to guide decisions about when to stop searching, although to our knowledge they have not been used in CEE evidence-synthesis to date (reviewed in Glanville, in press).

For web-searches (e.g. using Google) it is difficult to provide specific guidance on how much searching effort is acceptable. In some evidence syntheses, authors have chosen a "first 50 hits" approach (hits meaning articles, e.g. Smart & Burling 2001) or a 'first 200 hits' approach (Ojanen et al. 2014), but the CEE does not encourage such arbitrary cut-offs. What should be reported is whether stopping the screening after the first 50 (or more) retrieved articles is justified by a decline in the relevance of new articles. As long as relevant articles are being identified, the project team should ideally keep on screening the list of results.

3.3 Planning study eligibility criteria and eligibility screening

3.3.1 The Eligibility Criteria

Rationale for eligibility criteria

The use of pre-specified and explicit eligibility criteria ensures that the inclusion or exclusion of articles or studies from a systematic review or systematic map is done in a transparent manner, and as objectively as possible. This reduces the risk of introducing errors or bias which could occur if decisions on inclusion or exclusion are selective, subjective, or inconsistent. An objective and transparent approach also helps to ensure reproducibility of eligibility screening. Failing to consistently apply eligibility criteria, or using criteria which are not relevant to the evidence synthesis question, can lead to inconsistent conclusions from different evidence syntheses (e.g. illustrated by Englund et al. 1999 for stream predation experiments and McDonagh et al. 2014 for health research studies).

The eligibility criteria for a systematic review or systematic map should reflect the question being asked and therefore follow logically from the 'key elements' that describe the question structure. Many environmental questions are of the 'PICO' type, where the interest is on determining effects of an intervention within a specified population. For a PICO-type question the key elements (P, I, C, O) would specify which population(s), intervention(s), comparator(s) and outcome(s) must be reported in an article describing a primary research study in order for

that article to be eligible for inclusion in the evidence synthesis (examples of PICO and other types of question structure are given by EFSA, 2010; Aiassa et al., 2016; and James et al., 2016).

Developing your search strategy can in turn help define or refine eligibility criteria that will be used for the screening of the literature once the full search is conducted (see Section 6). Titles and abstracts and full text found during scoping can form a sample of the literature within which papers that are not relevant (ineligible) for different reasons (including unexpected use of synonyms, or use of similar wording in other disciplines) may be identified and appropriate eligibility criteria developed. Planning eligibility criteria allows for discussion with the commissioner about the scope and scale of the articles that will be retained and the finalised eligibility criteria will be reported later on in the evidence synthesis Protocol.

An example of eligibility criteria for an environmental intervention (i.e. PICO-type) systematic review question is shown in Box 3.2, for the question 'What are the environmental and socioeconomic effects of China's Conversion of Cropland to Forest Programme (CCFP) after the first 15 years of implementation?' (Rodríguez et al. 2016). As the example illustrates, eligibility criteria may be expressed as inclusion criteria and, if helpful, also as exclusion criteria.

Box 3.2 Example systematic review eligibility criteria in relation to question key elements for an intervention (PICO-type) environmental systematic review question (from Rodríguez et al., 2016)

Question: "What are the environmental and socioeconomic effects of China's Conversion of Cropland to Forest Programme (CCFP) after the first 15 years of implementation?"	
Question key elements	Eligibility criteria
Populations (P): • CCFP enrolled lands (cropland/ wasteland/ ecological trees/ economic trees) • CCFP households and their individual members	Included: Both human populations and land resources, including CCFP participant households, their individual members and their CCFP enrolled lands (cropland, wasteland, ecological trees, and economic trees). Excluded: Grasslands, since they no longer form part of the CCFP and because they contribute to significantly different environmental outcomes as compared with forests.
Interventions (I): • CCFP (subsidies, skill-training, and enforcement with field checks)	Included: CCFP compensation subsidies, skill training for local farmers, and enforcement work with field checks, and all information on other types of subsidies that might have an impact on household livelihoods and the environment. Excluded: Natural Forest Protection Programme, as this does not overlap with the CCFP.
Comparators (C): • Non-enrolled sloping lands, and lands prior to CCFP implementation • Non-participant households, and households prior to CCFP implementation	Included: Non-enrolled sloping lands, and lands prior to CCFP implementation; and non-participant households, and households prior to CCFP implementation. Included 'before-and-after' comparators in both human populations (i.e. the socioeconomic status of both participant and non-participant households before and after the CCFP interventions) and land resources (i.e. the environmental status of both enrolled and non-enrolled lands before and after the CCFP intervention).
Outcomes (O): • Environmental outcomes (changes in water discharge, soil erosion, flood risk, local biodiversity, etc.) • Socioeconomic outcomes (changes in household income structure, migration, etc.)	Included: Soil erosion and flood prevention, reconversion of forestland to cropland, land-use and forest cover change, tree survival rates, biomass and carbon storage, and biodiversity. Income, employment, food security, land access and social equality, and migration. Excluded: Studies assessing potential or future outcomes of the CCFP, including model projections or other predictions of program impact, as the review only sought to assess the actual impacts of CCFP implementation (i.e. those which have already taken place).
Note: Study design eligibility criteria were also specified by Rodríguez et al. (2016); for brevity these are not reproduced here.	

Ideally, the eligibility criteria should be specified in such a way that they are easy to interpret and apply by the review team with minimal disagreement. For some systematic review or systematic map questions the eligibility criteria may be very similar to or identical to the question key elements and the question itself, whereas in other cases the eligibility criteria may need to be more specific, to provide adequate information for the review team to make selection decisions.

In the example systematic review question (Box 3.2) it is clear that if an article describing a primary research study did not provide information on the intervention (i.e. the Conversion of Cropland to Forest Programme) then it would not be appropriate for answering the review question. As such, the article could be excluded. Similarly, an article that did not report any environmental or socioeconomic outcomes would not be relevant and could be excluded. The example question illustrates that articles can be efficiently excluded if they fail to meet one or more inclusion criteria; they are included only if they meet all the eligibility criteria.

Keeping the list of eligibility criteria short and explicit, and specifying the criteria such that an article would be excluded if it fails one or more of the criteria is a useful approach since this minimises the range of information that members of the review team would need to locate in an article and means that if an article is clearly seen not to meet one of the criteria then the remaining criteria would not have to be considered. Since a single failed eligibility criterion is sufficient for an article to be excluded from an evidence synthesis, it may be helpful to assess the eligibility criteria in order of importance (or ease of finding them within articles), so that the first 'no' response can be used as the primary reason for exclusion of the study, and the remaining criteria need not be assessed (Higgins & Green, 2011).

The example in Box 3.2 is for a relatively broad systematic review question. For a systematic map the question may be even broader since the objective of a map is to provide a descriptive output. Irrespective of how broad the question is, the process for developing eligibility criteria which we have outlined here applies both to systematic reviews and systematic maps (James et al., 2016).

Study design as an eligibility criterion

The types of primary research study design (e.g. observational or experimental; controlled or uncontrolled) that can answer an evidence synthesis question will vary according to the type of question. The study design is sometimes made explicit in the key elements (e.g. 'PICO'- type questions may be stated as 'PICOD' or 'PICOS' in the scientific literature, where 'D' (design) or 'S' (study) indicates that study design is being considered) (e.g. Rooney et al., 2014). Even if study design is not explicit in the question structure it should be considered as an eligibility criterion. This is particularly important for systematic reviews since the designs of studies that are included should be compatible with the planned approach for the data synthesis step (e.g. some meta-analysis methods may specifically require controlled studies). The type of study design may also be indicative of the likely validity of the evidence, since some study designs may be more prone to bias than others (see Box 3.3). Note that in systematic reviews the full assessment of risks of bias and other threats to validity takes place at the critical appraisal step, and this should always be conducted irrespective of whether any quality-related eligibility criteria have been specified.

Box 3.3 Overview of research designs

There are three major classifications of research designs and the types of data collected vary depending on the design: observational research, correlational research, and experimental research. In observational research, researchers can either observe the target population (human and non-human without attempting to influence it, or compare cases that received the 'treatment' (through some unknown selection mechanism) with those that did not receive it. In correlational research, researchers examine the covariation of two or more variables, but without distinguishing which variable is affecting others. In experimental research, researchers control conditions to determine how one variable affects others. The target population is assigned to groups, preferably randomly. There is a control group that is not subject to the variable being examined and at least one experimental group that is subject to the variable. The groups are compared to examine the effect of the variable being investigated. Laboratory experimentation epitomises this approach, but is not always of relevance to questions concerning environmental management. Field experimentation attempts to simulate as closely as possible the conditions under which a causal process occurs, the aim being to enhance the external validity, or generalizability, of experimental findings (Gerber and Green, 2011).

Studies can use quantitative data, qualitative data, or both types of data (mixed methods). Each approach has advantages and disadvantages. A full discussion of data types is outside the scope of these Guidelines; useful references that discuss data types in detail include, for example, Bernard (2006). Briefly data are 'quantitative' if they are in numerical form and 'qualitative' if they are not. Qualitative data can include more than text: photographs imagery and sound recordings are widely used in primary environmental research (e.g. Gibson and Russell 2006, Beason 2014). Debate about the relative merits of these two data types can be unhelpful, obscuring the fact that qualitative and quantitative data are closely related to each other. All quantitative data is based upon qualitative judgments; and all qualitative data can be described and manipulated numerically. For the purposes of evidence synthesis, it is important to understand what types of data are reported in the primary research and to decide which types will be used in the synthesis.

Quantitative variables can be continuous or discrete.

- Continuous: the variable can, in theory, be any value within a certain range. Can be measured (e.g. area of study site, height of tree)
- Discrete: the variable can only have certain values, usually whole numbers, which can be counted (e.g. number of individuals completing a questionnaire, number of times a field has been ploughed in the past 20 years)

Qualitative variables can be nominal or ordinal.

- Nominal: the variable does not have a specific order (e.g. biome type, farming system, ethnicity)
- Ordinal: the variable has a specific order. (e.g. the DAFOR (Dominant, Abundant, Frequent, Occasional, Rare) scale, IUCN Threat Levels.

3.3.2 Pilot testing the eligibility criteria and screening process

The eligibility screening procedure should be pilot-tested and refined by arranging for several reviewers (at least two per article) to apply the agreed study inclusion (eligibility) criteria to the subset of identified relevant articles. A typical approach is to develop an eligibility screening form that lists the inclusion and exclusion criteria, together with instructions for the reviewers, to ensure that each reviewer follows the same procedure. A standard approach is to develop a form that guides the reviewers to make simple decisions, for example: to include the article; to exclude it; or to mark it as unclear. Reviewers screen the titles and/or abstracts of the subset of articles and then compare their screening decisions to identify whether they are adequately consistent. If necessary, the form should be refined and re-tested until an acceptable level of agreement is reached. Once the suitability of the eligibility form has been tested on titles and/or abstracts, it should be tested on full-text versions of articles in the identified subset using a similar approach. The finally agreed draft eligibility screening criteria and form should then be provided when the Protocol is submitted (see below).

Pilot testing is important for validating reproducibility and reliability of the method. Pilot testing can:

- check that the eligibility criteria correctly classify studies;
- provide an indication of how long the screening process takes, thereby assisting with planning the full evidence synthesis;
- enable agreement between screeners to be checked; if agreement is poor this should lead to a revision of the eligibility criteria or the instructions for applying them;
- provide training for the review team in how to interpret and apply the eligibility criteria, to ensure consistency of understanding and application;
- identify unanticipated issues and enable these to be dealt with before the methods are finalised.

The eligibility screening process should be tested on a sample of articles. There is no firm 'rule' about how many articles should be tested, but the review team will need to satisfy themselves that the eligibility criteria will correctly identify articles that can answer the evidence synthesis question without needing any further amendments. Higgins & Green (2011) suggested using around 10-12 articles, including ones which are thought by one screener to be definitely eligible, definitely ineligible, and doubtful, and can be screened by one or more further members of the review team to assess consistency. Pilot testing should be performed for each separate step of the screening process that will be conducted, i.e. the title, abstract (or title plus abstract) and full-text screening steps.

If relevant articles are found to have been excluded, irrelevant articles are included, or a large number of 'unclear' judgements are being made by the review team, then the eligibility criteria should be revised and re-tested until an acceptable discrimination between relevant and irrelevant articles is achieved. The finally-agreed eligibility criteria should then be specified in the evidence synthesis Protocol.

3.4 Planning for data coding (Systematic Reviews and Maps) and data extraction (Systematic Reviews)

Data coding and data extraction refer to the process of systematically extracting relevant information from the articles included in the Evidence Synthesis. Data coding is the recording of relevant characteristics (meta-data) of the study such as when and where the study was conducted and by whom, as well as aspects of the study design and conduct. Data coding is undertaken in both Systematic Reviews and Systematic Maps. Data extraction refers to the recording of the results of the study (e.g. in terms of effect size means and variances). Data extraction is undertaken in Systematic Reviews only. A standard data coding or extraction form or table (e.g. spreadsheet) is usually developed and pilot-tested on full-text copies of the relevant subset of identified articles. The table contains prompts to the reviewers to record all relevant information necessary to address the synthesis question, plus any additional information required for critical appraisal (see below) and any contextual information that will be required when writing the final Evidence Synthesis report. As with the eligibility screening step, the pilot test should involve at least two reviewers per article, so that any inconsistencies can be identified and corrected. Any issues with data presentation should be noted at this point, so that they may inform synthesis planning. For example, Review Teams may find that data are not consistently presented in a suitable format and that they may need to contact original authors for missing or raw data. The finally agreed draft data coding or extraction table should then be provided when the Evidence Synthesis Protocol is submitted (see Section 4). Data coding and extraction tables for Systematic Reviews are likely to be more detailed than Data coding tables for Systematic Maps, reflecting the different principles of these Evidence Synthesis methods (as explained in Section 2). Data coding in Systematic Reviews should take into account capture of information on potential reasons (effect modifiers) for heterogeneity in outcomes.

3.5 Developing Critical appraisal criteria (Systematic Reviews only)

3.5.1 Why is critical appraisal necessary?

Not all research is conducted to the best standards of scientific rigour and therefore not all information available about a particular topic may be correct. A key challenge is to identify information which is likely to be correct and that which is not. If a systematic review is based on incorrect evidence then the results of the review will also be incorrect. The critical appraisal step is a crucial part of a systematic review since this is where the "correctness" of the evidence is ascertained and decisions are made as to which evidence is permitted to inform the review's conclusions. For this process to work effectively, two key criteria have to be met: First, the critical appraisal should focus on aspects of research study conduct that influence whether the resulting information will be correct or not (bearing in mind that some aspects of study design may be more important than others). Second, to have any bearing on the review's conclusions the critical appraisal step has to directly inform the data synthesis step of the systematic review. It is implicit from this that critical appraisal should be not only a structured process but one that has to be planned a priori. As with the other key steps of a systematic review, the methods of critical appraisal should be pre-specified in the review Protocol.

Critical appraisal refers to the process of assessing whether the evidence is valid for answering the Review question. **Key aspects of validity are "internal validity" which is the extent to**

which evidence is free from bias or confounding, and "external validity" which is the extent to which the evidence is relevant to the question being asked (i.e. whether it can be generalised from the original study to address the review question). Other aspects of evidence "quality" can also be assessed if considered important. The critical appraisal process requires reviewers to use pre-specified criteria to make judgements about whether validity and other quality criteria are met (often "yes", "no" or "unclear" judgements). Review teams are advised to use the *CEE Critical Appraisal Tool* to assist in this process.

In developing a checklist using the tool, review teams may find it useful to think of a theoretical gold standard methodology that a primary study might adopt to minimise bias and maximise analytical power. The gold standard may be practically impossible but nevertheless possible to describe in theory. The checklist can then be based on the impact of elements of the gold standard being missing (e.g. measurements at baseline or randomization). Ideally, the type of bias (see below) that each missing element potentially introduces should be listed.

This checklist should be pilot tested on the full-text version of each article in the sample of potentially relevant references, by at least two reviewers per article. Reviewers can then compare their judgements and inconsistencies or disagreements can be taken into account when improving the critical appraisal process and checklist. The finally agreed draft critical appraisal checklist should then be provided in the Evidence Synthesis Protocol (see Section 4).

3.5.2 Internal validity: Understanding bias

Bias is defined as a systematic deviation in study results from their true value, i.e. it means either an underestimation or overestimation of the true value. The magnitude of bias can range from trivial to substantial. Bias should not be confused with statistical uncertainty as a result of random error, which is present in all research studies. Random error reflects inaccuracy of estimation that is distributed randomly around the true result. Often, random error can be reduced by increasing the sample size in a research study, or by quantitatively combining the results of similar studies in a meta-analysis (subject to the studies being adequately comparable), hence improving the precision of the result (Glass, 1976). Bias, on the other hand, refers to a systematic error which cannot be reduced by increasing the sample size or by pooling study results in a meta-analysis. If bias is present in primary research studies their results will be incorrect. It is generally acknowledged that bias is an important threat to the validity of research findings across scientific disciplines, and it has been argued that bias is one of several factors that collectively contribute to the majority of research findings being incorrect (Ioannidis, 2005). Traditional non-systematic reviews of evidence which do not formally assess the rigour of primary research studies would not be able to detect bias.

A misleading result from an evidence synthesis could occur where a precise but wrong answer is generated (e.g. a point estimate that is incorrect but has a narrow confidence interval). This could arise for example if the included studies in a meta-analysis exhibit consistent systematic error with relatively low random error (Figure 3.2). To avoid this kind of misleading result, it is clearly important that risks of bias are sought and if possible identified before the data synthesis step of a systematic review takes place.

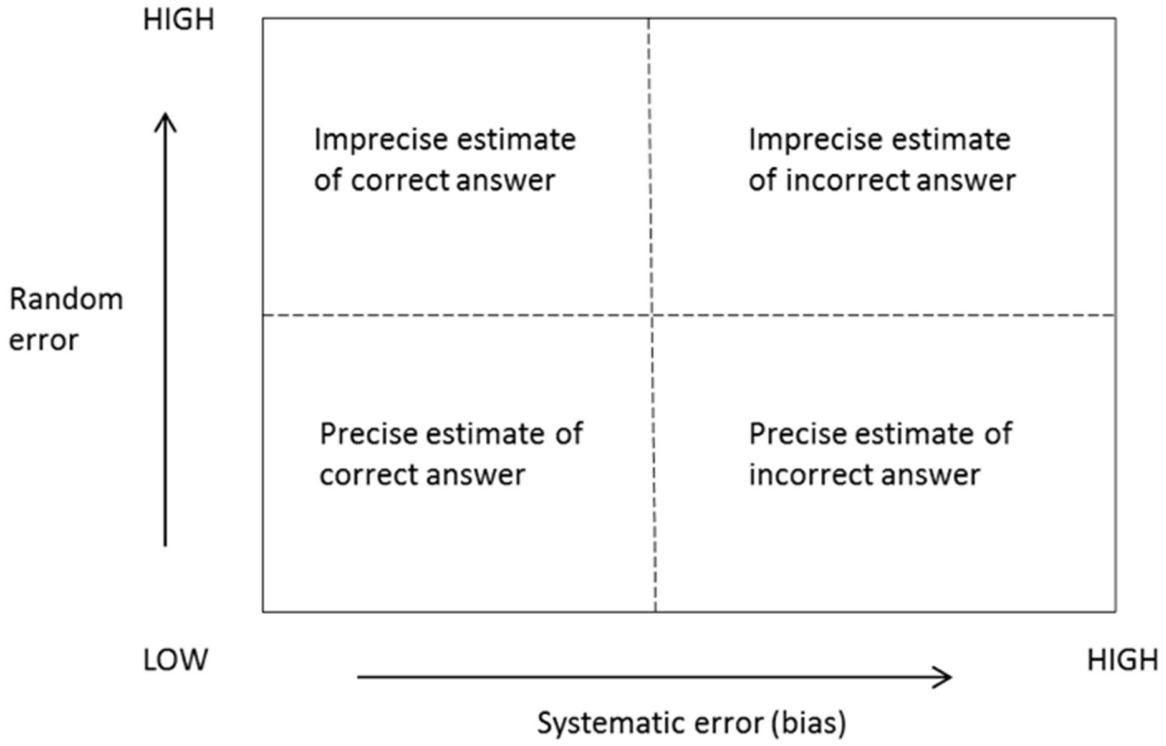


Figure 3.2. Schematic illustration of the potential influence of random error and systematic error (bias) on a study outcome

Bias in research studies can arise for a variety of reasons. Poor design of a research study may mean that it consistently underestimates or overestimates the true value of an outcome and the study researchers may not be aware of this. In some cases researchers may have a vested interest in a particular outcome and this could lead, either intentionally or unintentionally, to various types of bias. Considerable experience from evidence synthesis in health research has shown that where bias is present it often leads to over-estimation of beneficial outcomes, e.g. exaggerating the actual benefits of an intervention such as a drug treatment (Higgins et al., 2011).

The concept of "risk of bias"

Evidence for the existence of bias comes from meta-epidemiological health research that has assessed large numbers of studies to determine whether outcomes differ systematically between studies that have a particular design feature and those that do not (e.g. Wortman 1994; Schulz et al. 1995; Chan et al. 2004; Wood et al. 2008; Liberati et al. 2009; Kirkham et al. 2010; Higgins & Green, 2011; Holman et al. 2015). However, it is usually impossible to directly measure bias within individual primary research studies. Instead, an indirect approach is to infer the "risk of bias" by examining the study design and methods to determine whether adequate steps were taken to protect against bias. Studies that fail to meet specified criteria for mitigating known types of bias may be referred to as being at "high risk of bias" whilst studies with adequate methodology to protect against bias are considered to be at "low risk of bias" (Higgins et al. 2011).

Meta-epidemiological studies on randomised controlled trials of interventions in health research have identified five main types of bias that the trials need to protect against to ensure that their results would be unbiased. These are selection bias, performance bias, detection bias, attrition bias, and reporting bias. These and other types of bias are explained in more detail later in this section. To understand and be able to identify the different types of bias that may arise in research studies, Review Teams should be familiar with the concepts of confounding and effect modification.

Confounding and effect modification

To assess whether there might be a risk of bias, it is important to understand the interrelationships between the explanatory variables and dependent variables that are present in a study. In a well-conducted research study of causation the intervention/exposure would be the explanatory variable, the effect (and hence the measured outcome) would be the dependent variable, and these would be linked by one or more clearly-specified cause-effect pathways. Numerous other variables, which could act as covariables in relation to the study hypothesis, are likely to be present in the system under study and these would need to be controlled for in the study design to ensure that inferences based on the measured outcomes accurately reflect the hypothesised effect of the intervention/exposure. The variables are often categorised as prognostic variables, effect modifiers and confounding variables (or confounders) in the evidence synthesis literature, although sometimes these may be referred to by a variety of synonyms (Peat 2001).

- A prognostic variable is a variable that is known (e.g. based on knowledge from previous empirical research), or considered very likely (e.g. based on plausibility and probability) to predict the outcome of interest.
- An effect modifier is a variable which differentially (positively or negatively) influences an outcome by interacting with the cause-effect pathway, but is not a causal factor in itself (i.e. it does not modify the intervention/exposure). The observed cause-effect association will be correct in principle, but the outcome will be biased (systematically under-or over-estimated) if the effect modifier is not controlled for in the study. An effect modifier may also be a prognostic variable.
- A confounder is a variable external to the cause-effect pathway that interacts with both the intervention/exposure and the outcome. A confounder would meet these three criteria: (1) it is a predictor of the outcome, independent of the intervention/exposure; (2) it is associated with the intervention/exposure; and (3) it is not in the causal pathway between the intervention/exposure and outcome. Presence of a confounder means that the observed cause-effect association is not correct and so the outcome will be biased if the confounder is not controlled for in the study.

Figure 3.3 provides a schematic summary of how these types of variable interact with the intervention/exposure and outcome of interest.

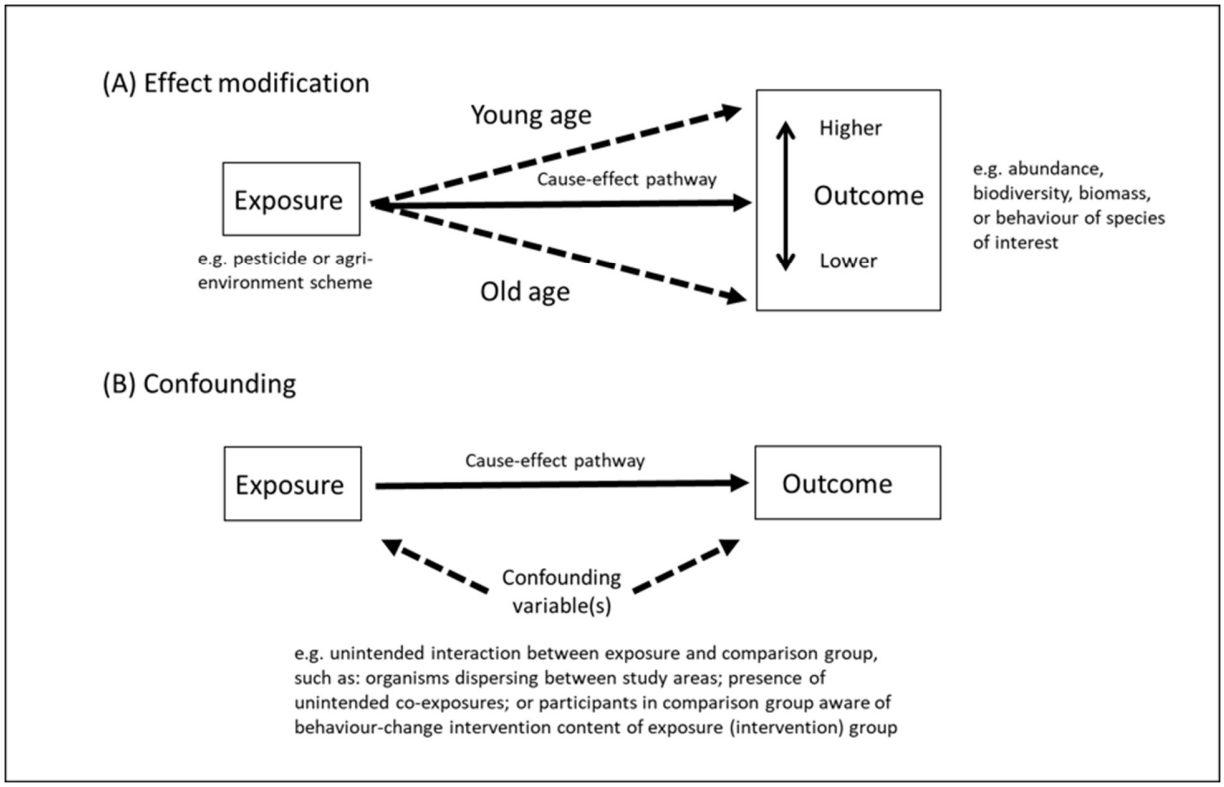


Figure 3.3. Schematic illustration of interactions to look for when investigating potential sources of bias in research studies

Whether a variable is a prognostic variable, effect modifier, and/or confounder will depend on the outcome and exposure being assessed. If a study is asking whether a pesticide influences the fecundity of an organism, then age would almost certainly be a prognostic variable since it is known from empirical research across a wide range of organisms that fecundity varies strongly with age.

A prognostic variable can be defined in isolation of any intervention/exposure (i.e. the prognostic influence of age upon fecundity does not require there to be an intervention/exposure). An effect modifier on the other hand can only be defined in the context of a putative effect of interest, meaning that a putative cause-effect pathway for an intervention/exposure would need to have been specified. As such, effect modifiers are treatment-specific. Supposing that the effect that a pesticide has on fecundity varies with an organism's age, then age would be both a prognostic variable and an effect modifier.

The term "confounder", or "confounding variable", is sometimes used in the scientific literature in a general sense to mean any covariate that could predict the intervention/exposure or outcome (i.e. referring to both confounders and effect modifiers as defined above). However, in statistical analysis it is important to distinguish between confounding variables and effect modifiers. This is because confounders exhibit collinearity with both the intervention/exposure and outcome whilst effect modifiers exhibit collinearity with the outcome but not the intervention/exposure. A challenge in environmental research studies is to understand which of the many biotic and abiotic

variables and their interactions that are present in ecological systems could be confounding variables or effect modifiers in relation to the study hypothesis. A conceptual model can be a helpful means of visualising the key variables that relate to the intervention/exposure and outcome, so as to clarify which may be confounders or effect modifiers.

The relationships shown in Figure 3.3 highlight the types of variables and interactions that review teams should look for in the system to which the review question relates and may form a useful basis for developing a conceptual model to help ensure that key variables and interactions have not been missed.

Principles for assessing risk of bias

Extensive experience of conducting critical appraisal of studies in systematic reviews of health topics has identified several core principles that should guide how risk of bias is assessed (e.g. Higgins et al. 2011):

Assessment should focus on internal validity

Internal validity indicates whether the results are correct or not (i.e. biased). This should be distinguished from random error (precision), external validity (i.e. generalisability), and quality of reporting, which do not themselves indicate whether bias is present. Note that some aspects of study "quality", such as whether sample size was calculated, are not related directly to the risk of bias (Higgins et al. 2011). Critical appraisal assessments which mix up these different aspects of study "quality" or reporting would not be able to clearly detect threats to internal validity. External validity, which is explained below, should be assessed separately from internal validity.

Risk of bias should be assessed separately for each outcome

Risks of bias are likely to differ according to the result being assessed (Page & Higgins 2016) and should therefore be assessed separately for each outcome rather than for the study as a whole (unless it can be justified that outcomes are similar enough that they would be subject to the same risks of bias).

3.5.3 Criteria for identifying risks of bias in environmental research studies

In this section we provide lists of scenarios that can help to identify risks of bias in each of the core domains, i.e. selection bias (Box 3.4), performance bias (Box 3.5), Detection bias (Box 3.6), attrition bias (Box 3.7) and reporting bias (Box 3.8). Where possible, we have contextualised the scenarios with actual or hypothetical examples from environmental management research (see the descriptive text below for each domain of bias). Unless stated otherwise, the scenarios are likely to be broadly applicable across a range of study designs.

The aim of this section is to guide review teams on where to look for risks of bias in environmental management studies, but the lists of scenarios are not exhaustive. Review teams should check whether further confounding variables or effect modifiers are present in addition to

those listed. The identification of risks of bias is an iterative process and pilot testing of the process is essential to enable the review team to become adept at identifying risks of bias.

Selection bias (Box 3.4)

Selection bias is an inherent concern in all types of study design and can only be controlled by ensuring that the study units (e.g. people or animals being assigned to intervention or comparator groups or chemicals being assigned to plots in a field experiment) are allocated randomly. A key challenge that review teams face when looking for similarity among study groups is to know which of the factors might or might not be potential confounding variables or effect modifiers, although sometimes these may be obvious. For instance, in human and animal studies age and health status are very likely to be effect modifiers (i.e. they are likely to systematically influence the outcome if not balanced between study groups). In agricultural field experiments soil type is very likely to be a confounding variable or effect modifier given that it is a key determinant of biotic and abiotic diversity (soil type is correlated with other factors such as geographical location and vegetation type and so these factors would also likely be confounders or effect modifiers).

Key issues to look for are:

- Lack of randomisation (i.e. no randomisation, or randomisation is stated but not appropriately implemented). All types of study that lack random allocation are inherently at risk of selection bias since unmeasured confounding variables cannot be controlled for.
- In randomised studies: Study investigators may be able to influence the allocation process, preventing it from being truly random (e.g. preferentially selecting which participants are assigned to each study group, or which interventions or exposures are assigned to study plots or areas). Concealment of the intervention or exposure allocation should always be feasible in well-conducted studies (though may not be commonly implemented in environmental research).

Non-randomised studies are inherently prone to selection bias, but steps can be taken in some types of study design to control for selection bias as much as possible. These include selecting populations (or study areas) that appear to be as similar as possible such that the comparator (or control) group is sampled from the same population as the intervention (or cases) group, and/or using statistical correction to ensure that the groups are matched on all the known variables that could influence outcomes. This cannot account for any imbalances in unmeasured variables, but study investigators may make a pragmatic assumption that the measured variables are likely to be the most important confounders or effect modifiers.

A common problem in environmental research studies is that the baseline characteristics of the populations or study areas of interest are not always reported, which may preclude an assessment of the comparability of study groups. For example, Stewart, Coles & Pullin (2005) found that baseline data to confirm whether study sites were homogeneous before a vegetation burning intervention was applied were generally lacking. And Mackenzie Ross et al. (2016) found that several studies assessing neurotoxicity of low level exposure of people to organophosphate

insecticides did not provide any information on prior exposure before the study, which could be a confounding factor. This illustrates the importance of considering not only the characteristics of study groups at the start of a study but also any historical differences between groups that could introduce selection bias. If review scoping suggests that studies are likely to generally be deficient in reporting baseline information then review teams should consider whether it would be feasible to contact study authors for this information.

Box 3.4 Scenarios indicative of risk of selection bias

Selection bias: systematic error in an outcome caused by confounding as a result of baseline imbalances in characteristics of study units (e.g. participants or samples)

Key question: Are the intervention/exposure and control groups comparable on all important potential effect modifiers? Yes=low risk of bias; No=high risk of bias

Scenarios indicative of a high risk of selection bias:

- Study units are not allocated randomly (e.g. people or animals are not randomly allocated to exposure/control groups or farming interventions are not randomly allocated to plots in a field experiment)
- In randomised studies only: Randomisation is claimed, but was not appropriately implemented.
- In randomised studies only: Study investigators are likely to have influenced the allocation process (due to a lack of allocation concealment), preventing allocation from being truly random (e.g. if investigators can preferentially select which participants are assigned to each study group, or which interventions or exposures are assigned to study plots or study areas).
- In non-randomised concurrently-controlled studies: the comparator (or control) group was not sampled from the same population as the intervention/exposure (or cases) group, and the groups are not matched on all the known variables that could influence outcomes. For case-control studies, controls should represent the population from which cases arose (i.e. controls should have met the case definition if they had the outcome). Non-randomised studies should be considered by default to have high risk of bias unless the allocation process or statistical matching of groups can be justified to have accounted for all important potential effect modifiers.
- In case-control, cohort and cross-sectional studies: Criteria for including/excluding study populations or samples vary across the comparison groups.
- In observational studies with single groups, non-concurrent controls or historical controls: these study designs are inherently at high risk of selection bias and it is not possible to assess the risk of selection bias for single-group studies. The default judgement for these types of studies is likely to be a high risk of bias. However, selection bias may not be a meaningful type of bias to assess for some outcomes (e.g. if the interest is on detecting rare events, or where only a single time series of data is available), and in some cases it may be considered more meaningful to focus on external validity.

Depending upon the study design, imbalances in study groups may be quite subtle to detect. An example is provided by Duffy et al. (2014) in which the use of a standard test system would

result in unnaturally healthy controls in ecotoxicological testing of pharmaceutical effects on fish.

Note that the allocation of study groups can have implications both for selection bias and external validity (see Section 3.5.4). For example, if a cross-sectional study sampled a range of geographical sites there could be a risk of selection bias if the sites were not selected randomly, but also a threat to external validity if the randomly-selected sites were only a subset of those relevant to the review question.

Performance bias (Box 3.5)

Performance bias is a systematic error in the effect attributed to an intervention or exposure caused by the influence of a confounding factor. Performance bias may arise for several reasons, which may or may not occur together in the same study.

Study investigators who are aware of the allocations (e.g. of people or animals to groups, or crop treatments to plots in a field study) may be prone to inconsistency in how they manage the study groups, potentially favouring one group over the other (e.g. by being more meticulous in their adherence to the Protocol for one group). These "observer biases" are strongest when researchers expect a particular result, are measuring subjective variables, and have an incentive to produce data that confirm predictions. For example, students who believed their test rats had been selectively bred for maze-solving ability recorded better maze performance than did students told their rats were bred for poor maze-solving ability, despite both groups possessing randomly assigned, normal rats. This type of performance bias can be prevented by blinding study investigators to the group allocations. Although it is known that non-blind studies tend to report higher effect sizes and more statistically significant results, blinding is uncommon in the life sciences. It is not always feasible to blind researchers in environmental management studies. For example, where vegetation characteristics are likely to differ between study plots or areas (as with agri-environment or vegetation control interventions) the plant species composition or density would likely indicate the intervention that was allocated.

In environmental field studies performance bias may relate to the scale of the study. For example, in studies with insecticides the use of small plots can lead to an overestimation of the recolonization rate of invertebrates (Bero et al. 2016).

Box 3.5 Scenarios indicative of risk of performance bias

Performance bias: systematic error in an outcome caused by factors which modify the cause-effect relationship between an intervention/exposure and outcome

Key question: Are there any factors, other than the intended intervention or exposure, that may have influenced the outcome? Yes=high risk of bias; No=low risk of bias

Scenarios indicative of a high risk of performance bias:

- Study investigators are aware of (i.e. not blinded to) the allocation of intervention/exposure or control groups and as a result could manage the intervention/exposure or control groups differently.
- Co-interventions or co-exposures are present that are not part of the intervention or exposure being assessed, and these differ (in type or amount) between the intervention/exposure and control groups.
- There are other factors, interactions or processes not accounted for in the study design that could link interventions or exposures with outcomes and may therefore act as confounders/effect modifiers
- Contamination of study groups occurs (e.g. control group participants had knowledge of intervention details; or organisms moved between different intervention or exposure groups).
- In before-after studies or time series studies: One or more background secular trends are likely to influence the study outcome.
- Study participants (i.e. the population under study) are aware that they are in a research study or a specific study group and this knowledge could influence how they respond to an intervention or exposure.
- There are deviations from the study protocol

Detection bias (Box 3.6)

Detection bias may arise if there are systematic differences in the way outcomes are assessed among the study groups being compared. Possible sources of detection bias are: systematic misclassification of the exposure, intervention or outcome (e.g. because of variable definitions or timings of assessments), inconsistent application of diagnostic thresholds across study groups, the need for recall from memory (e.g. in surveys or questionnaires), inadequate assessor blinding (such that the investigator's knowledge of the study group allocations could influence how they measure and/or record outcomes), and faulty measurement techniques.

Where organisms are being sampled in their natural environment, detection bias might arise if there is a systematic difference between study groups in the investigators who are assigned to do the sampling. For example, bias might be introduced if one group of investigators always sampled the exposure plots whilst a different group of investigators always sampled the comparator plots; or if the investigators assigned to sampling the exposure plots had different training in sampling, or other relevant expertise, compared to those who sampled the comparator plots. Random assignment of outcome assessors to study groups, and blinding of the outcome

assessors so that they are unaware of the group allocations, are ways to reduce the risk of detection bias although, as mentioned above, blinding is uncommon in the life sciences.

Sampling devices could introduce bias if their capture efficiency is variable and differs systematically between study groups. It is well-known, for example, that the capture efficiency of pitfall traps and suction samplers for sampling terrestrial invertebrates is dependent upon vegetation characteristics and habitat structure. Another issue with pitfall traps is that they depend upon organisms' activity (which is related to temperature and body size) and therefore they provide a measure of "activity-abundance" rather than an estimate of abundance. If studies using pitfall traps are claimed to be providing abundance estimates without accounting for between-group differences in activity then bias could be introduced. Usually, a range of methods is available for sampling organisms, e.g. for reptiles or invertebrates, and these can differ in which taxa they sample, so it is important that the review team is experienced enough to know which sampling methods would be most appropriate for the answering the review question without introducing bias. Given that many environmental management studies will involve manipulations of vegetation or habitat structure (e.g. studies involving herbicides, fertilisers, comparisons of crops, agri-environment schemes, or other environmental management prescriptions such as burning or drainage), there is considerable scope in research studies for sampling efficiency to be confounded with these factors if they differ systematically between study groups and are not accounted for in the study design.

Box 3.6 Scenarios indicative of risk of detection bias

Detection bias: Systematic error in an outcome caused by the way the outcome is measured

Key question: Was the outcome assessment method likely to have accurately measured the outcome, and was the method of outcome assessment the same across all relevant study groups? Yes=low risk of bias; No=high risk of bias

Scenarios indicative of a high risk of detection bias:

- Study investigators who were responsible for assessing outcomes are aware of (i.e. not blinded to) the allocation of intervention or exposure groups and the outcome assessments are such that the investigator could (intentionally or unintentionally) influence them.
- The timing of outcome measurements differs between intervention or exposure and control groups.
- A confounding factor links the sampling method to both the intervention/exposure and the outcome (e.g. if vegetation density differed between intervention/exposure and control groups and also influenced the efficiency of a sampling method for sampling birds, mammals or butterflies).
- The sampling method or measuring approach differs between study groups (e.g. if suction traps used in different study plots have different suction efficiency).
- The way an outcome is defined is not consistent between study groups, sampling times, or locations.
- The sampling method does not distinguish between the intended outcome measure and a related one that would likely have been influenced by the intervention or exposure (e.g. pitfall sampling of invertebrates where the catch is dependent on locomotor activity, or visual census of butterflies where the observation is dependent on flight activity, and where the activity is likely to be influenced (directly or indirectly) by the intervention or exposure.

Attrition bias (Box 3.7)

Attrition bias may occur where there are systematic differences between groups in the loss of participants, organisms, or samples from a study and the missing observations are related to the intervention/exposure and/or the outcome of interest (i.e. the missing observations are systematically different from those which remain in the analysis). Attrition bias can potentially change the collective (group) characteristics of the study groups and their observed outcomes in ways that affect study results by confounding and spurious associations. For example, if animals which die following exposure to a chemical are excluded from an analysis, this would create an imbalance between groups in the sensitivity of those animals that remain in the study (since the most sensitive animals have been excluded, those remaining in the analysis would be unrepresentatively insensitive in the exposure group). In cross-sectional studies such as surveys, non-response of participants could introduce systematic error if the reasons for non-response are related to the intervention/exposure or the outcome being assessed.

The risk of bias may be less, and perhaps could be considered trivial by the review team, if the proportion of missing data is small and/or the reasons for data being missing do not differ systematically between the study groups (e.g. if they can be assumed to be missing at random). However, clear justification should be provided for any assumptions made about missing data.

Box 3.7 Scenarios indicative of risk of attrition bias

Attrition bias: Systematic error in an outcome caused by missing data
Key question: If there were missing data were these of minor numerical importance (relative to the total sample size and effect estimate), balanced across the study groups and unrelated to the intervention/exposure and outcome? Yes=low risk of bias; No=high risk of bias Scenarios indicative of a high risk of attrition bias: • There is a large proportion of missing data and the number of missing samples or observations is unbalanced between study groups • Reasons for the data being missing are, or are likely to be, related to the intervention or exposure that is being assessed and/or the outcome being assessed. • The times at which data are missing differ between the study groups (e.g. data are missing earlier in one group than in the other, or the follow-up period differs between groups). • In randomised studies only: An intention-to-treat analysis is specified (all randomised participants are intended to be analysed in the groups to which they were originally randomised) but not all participants are included. • Study investigators' knowledge of group allocations could enable them to systematically influence which study participants or samples are excluded from the analysis (not applicable if there are no, or only minimal, missing data).

Reporting bias (Box 3.8)

Reporting bias refers to selective disclosure of results such that the outcomes that are reported do not provide a true reflection of the results that would have been observed had all measured outcomes been reported. In order to check for reporting bias the review team will need to have access to a statement of which outcomes were measured in the study. Ideally, this would be found in the study protocol. However, protocols are not commonly provided for environmental research studies and the review team may therefore need to consult the methods section (and possibly other sections) of the study report to ascertain which outcomes were measured.

Types of selective disclosure of results that review teams should be aware of are: reporting results for selected sampling times; reporting results for selected species or other taxa from among a wider list of taxa sampled (e.g. preferentially reporting the most or least sensitive

species to an exposure); reporting the most or least sensitive of a range of biomarkers or other outcomes measured; reporting incomplete data for outcomes (e.g. continuous data presented as categorical data with arbitrary cut-offs); and preferential reporting of only statistically significant (or statistically non-significant) results. Selective reporting could be a problem in studies that use multiple ways of assessing the same outcome but do not report all of these (e.g. if diversity is being measured using various different indices such as species richness, Shannon-Wiener, Simpson, and Berger-Parker indices), or in studies that employ multiple sampling methods but do not report results from all of them.

The review team should consider carefully whether non-reporting of outcomes would likely introduce bias, since there may be cases where non-disclosure of outcomes might be considered inconsequential or relatively unimportant. For example, non-reporting of short-term measurements in a long-term study may be considered less likely to misrepresent the true findings of the study than if the long-term measurements are not reported and only short-term measurements given. The review team should provide a clear rationale for their judgements made about the risk of reporting bias.

Box 3.8 Scenarios indicative of risk of reporting bias

Reporting bias: Systematic error in an outcome caused by selective (non-) reporting of results
Key question: Are all the measured outcomes fully reported; if not, are the missing data likely to be related to whether the results were positive or negative? Yes=low risk of bias; No=high risk of bias Scenarios indicative of a high risk of reporting bias: • No results are provided for an outcome that is stated to have been measured. • Partial results are provided for an outcome (e.g. for selected time points only). • Results are reported in an incomplete format (e.g. continuous data are reported as categorical data with arbitrary cut-offs). • Selective (non-)reporting of statistically significant outcomes or extreme values. • Incomplete reporting of the measured taxonomic or geographic resolution of the outcomes (e.g. results for some species or locations are missing, or only results for coarse taxonomic or functional units or areas are provided).

Other bias

"Other bias" refers to the presence of any further factors that could lead, directly or indirectly, to systematic underestimation or overestimation of outcomes or effect estimates but do not appear to be readily classifiable as one of the core bias types above. According to the original Cochrane Risk of Bias tool, the five core bias domains are independent. If additional confounding variables or effect modifiers are identified and are suggestive of a risk of bias then these should only be grouped under one of core domains in the recording template if they clearly relate to that domain. In cases of doubt as to whether an identified risk of bias can clearly be classified as

selection bias, performance bias, detection bias, attrition bias or reporting bias then it should be listed under "Other" risk of bias.

Bias arising from study sponsorship is an example of a type of bias that does not relate to any of the five core domains of bias is and would therefore be appropriate to list in the "Other bias" category. For example, in studies on non-human animals, non-industry sponsored studies were found to be more likely to conclude that the herbicide atrazine was harmful compared to industry sponsored studies (Duffield & Aebischer 1994).

3.5.4 External validity

External validity refers to whether the information obtained from a scientific research study is generalizable (i.e. directly applicable) to how the answer to the question being addressed would be applied in practice (Higgins et al., 2011). Experimental studies are typically conducted under controlled conditions that may not fully resemble those of the 'real world'. An experimental study of an intervention may demonstrate that the intervention can work under the specific conditions of the study, but we also need to know whether it would work in real-life field conditions where it is intended to be used. An intervention's performance in a study is termed "efficacy" whilst its performance in the real world is termed "effectiveness". Studies designed to reflect real-world conditions are referred to as "pragmatic" studies. External validity is important as it relates to how well efficacy predicts effectiveness (Khorsan & Crawford, 2014). Another element of external validity concerns whether the setting of a primary study included within a systematic review is appropriate to that of the review question being asked, for example whether the population, intervention, exposure, or outcomes in the primary study are comparable to those of the setting in which the answer to the review question is intended to be applied. Note that external validity is sometimes referred to in the literature as "generalisability", "applicability" or "directness".

The extent to which external validity should be assessed within a systematic review depends on whether the interest is on experimental or pragmatic studies and how the review question is framed (i.e. whether it is broad or narrow, and whether it captures efficacy and/or pragmatic studies). However, review teams should always consider two aspects of external validity: (1) whether the studies included in the review are appropriate for answering the review question; and (2) whether the answer to the review question can be applied directly by the intended end-user (which might, depending on the purpose of the review, be a conservation manager or other environmental practitioner; a policymaker; or a statistical model or process for which the review has generated a specified parameter).

The first aspect needs to be assessed for each individual primary study in the systematic review during the critical appraisal step of the review and the review team should specify in the Protocol the process that will be used if studies are judged to have low external validity (e.g. whether such studies would be excluded from data synthesis, or included in subgroup analyses or sensitivity analyses). The second aspect relates to how appropriate the review question is in relation to its intended purpose, and this should be considered during the development of the review question (rather than in the critical appraisal step of the systematic review).

3.5.5 Criteria for assessing the external validity of environmental research studies

There are two aspects of external validity: the extent to which the studies included in a systematic review are generalizable to answer the review question; and the extent to which the answer to the review question is generalizable to the setting in which the results of the review will be applied. The first of these is relevant to the critical appraisal step of a systematic review and the second is considered at the question development step.

The extent to which external validity of individual included studies will need to be assessed depends upon the breadth of the review's eligibility criteria and the nature of the included studies. For reviews with very narrow and clearly defined eligibility criteria it is unlikely that the studies included would lack external validity for answering the review question. However, sometimes it may not be clear how relevant studies are until they have been included and carefully scrutinised. This may be the case for studies on complex behavioural interventions for example, or studies that may be conducted at a range of different spatial and temporal scales.

A pragmatic way to assess the external validity of studies included in a systematic review is to consider systematically how well the key elements of the studies (e.g. PICO elements and study design) match those of the review question. Criteria that the review team should consider are: the relevance to the review question of the population; intervention/exposure; comparator; outcome; setting; geographical location; temporal scale; spatial scale; and study design.

3.6 Developing data synthesis methods (Systematic Reviews only)

Data synthesis refers to the collation of all relevant evidence identified in the Systematic Review in order to answer the review question. A narrative synthesis of the data should always be planned involving listing of eligible studies and tabulation of their key characteristics and outcomes. For Systematic Reviews, if evidence is available in a suitable format and quantity then a quantitative synthesis, such as aggregating by meta-analysis, may also be planned. The likely form of the data synthesis may be informed by the previous pilot-testing of data extraction and critical appraisal steps. For example, the Review Team may identify whether the studies reported in the articles are likely to be of sufficient quality to allow relatively robust statistical synthesis and what sorts of study designs are appropriate to include. This pilot-testing process should also inform the approach to the synthesis by allowing, for example: the identification of the range of data types and methodological approaches; the determination of appropriate effect size metrics and analytical approaches (e.g. meta-analysis or qualitative synthesis); and the identification of study covariates (see Section 9).

3.7 Estimating resource requirements

Whilst the process of scoping may seem like a time-consuming one, the benefits can be considerable and this early investment will allow the development of a comprehensive Protocol as well as improve the focus and efficiency of the review. Scoping should provide an estimate of the timeline of the review and team effort required so that a realistic budget can be prepared or the likely costs compared with the available resources.

Section 4

Writing and registering a Protocol

Last updated July 4th 2018

4.1 Purpose of the Protocol

CEE Evidence Syntheses require the publication of their Protocol (project plan) as an independent document, before the synthesis is conducted. There are several reasons for doing so:

- Data are not evidence unless accompanied by a Protocol and analysis.
- Within the CEE approach the Protocol acts as a formal registration of intent by the Review Team to conduct a CEE Evidence Synthesis on a given topic. It allows CEE to inform the scientific community of this project, and to let the Review Team know about any prior similar projects, or ongoing ones.
- The Protocol acts as an a priori guide and reference to the conduct of the synthesis that reflects views of stakeholders and that the Review Team and their commissioners agreed upon during the planning stage (including the scoping exercise).
- The Protocol is essential to minimise reviewer bias (e.g. resulting from ad-hoc decisions made or 'mission creep' during the synthesis process) and make the review process as rigorous, transparent, and well-defined as possible.
- The Protocol enables explicit and compulsory recording of any change that may occur during the conduct of the synthesis that would not have been foreseen. This is particularly important to ensure the confidence of the consumers, commissioners and stakeholders about the reasons for changes.

IMPORTANT: COMMITMENT TO REGISTER AND PUBLISH WITH CEE. By registering and publishing your Protocol with CEE you are registering your intent to conduct, and submit to for publication, a CEE Systematic Review/Map. You will be asked to confirm that you and your co-authors are aware of and agree with this commitment when you submit your Protocol for publication in Environmental Evidence.

4.2 Developing and writing a Protocol

The Protocol's background section should present the problem being addressed and the rationale for why a Systematic Review or Systematic Map is required. Where possible, a 'theory of change' or conceptual model should be presented that explains the process(es) whereby the intervention or exposure factor is thought to have an impact or cause a change in the subject population (see www.theoryofchange.org/what-is-theory-of-change/). In more complex situations a proposed causal chain, linking intervention(s) to outcome(s), may be helpful. The structure of an Evidence Synthesis Protocol mirrors the structure of the Systematic Review or Systematic Map that it guides. Beside a formal presentation of the question and its background (the "real world" context), a Protocol sets out (informed by the scoping process – see above) the

strategy for searching for relevant studies and defines eligibility criteria for article screening. The question elements defined in the question formulation stage provide the a priori inclusion criteria important for the objectivity and transparency of the synthesis. They should also lead to a description of the kinds of evidence (e.g. study designs) that you would consider valid to include in the synthesis. An Evidence Synthesis Protocol should also detail, with rationale, the likely methods to be used for eligibility screening, data coding/extraction, critical appraisal (Systematic Review only), and data synthesis, and state any conflicts of interest including details of any funding sources.

Since the Protocol sets out what the synthesis aims to achieve, it is useful for getting the engagement of experts who may have data to contribute. Anyone reading the Protocol should clearly understand the nature of the question and what type of evidence will inform it. Registering and posting of Protocols on the CEE website provides transparency and also acts as a record of which syntheses are in progress, enabling others to see if a synthesis is being conducted that may be of interest to them, or to prevent the initiation of a synthesis on a topic that is already underway. An example of Protocol development is given in Box 4.1. For examples of recently completed Protocols, visit the Environmental Evidence Library at: https://environmentalevidence.org/reviews-in-progress.

Once an Evidence Synthesis Protocol has been peer-reviewed and published as final, changes are discouraged. However, it may become necessary during the course of an evidence synthesis to make revisions because of deviations from the proposed methods. These changes should be clearly documented within the final synthesis report so that transparency and repeatability can be maintained. If a major change is necessary to a Protocol part-way through a Systematic Review or Systematic Map (e.g. change of question or major change in scope) then the Protocol should be updated in consultation with CEE, and the change should then be applied to all references, or studies, as appropriate, to avoid introducing bias. The final Evidence Synthesis report should explicitly state how the final Systematic Review or Map methods differed from the Protocol.

Protocols are plans of conduct and can rarely be fully comprehensive. They are judged in this context during the CEE peer review process. Consequently, the acceptance and publication by CEE of a synthesis Protocol does not guarantee acceptance of the resulting synthesis report. Problems with the latter may occur due to conduct that was not mentioned or not fully transparent in the Protocol.

As a general rule, a Protocol should set out the plan for a single report (a Systematic review or a Systematic map). Exceptionally, where there is a strong logical case made, a single Protocol may set out a plan for multiple reports. This should be anticipated and fully explained in the Protocol and should not be a post-hoc decision. Whether the multiple report route is permissible will be decided by the CEE Editorial Board and is a trade-off between the efficiency provided by the publication of one Protocol and the legitimacy and feasibility of combining several reviews or maps together.

Box 4.1 Example of Protocol Development

Following a suggestion from the Swedish Environmental Protection Agency, the MISTRA Council for Evidence-Based Environmental Management (EviEM) studied the feasibility of a Systematic Review on how mountain vegetation is affected by reindeer grazing. Review scoping was conducted, and the outcome was promising enough that the Swedish EPA remained committed to the idea. Using scoping as a basis, EviEM then drafted a first version of a review protocol. A Review Team was organised, and stakeholders (the Swedish EPA and other agencies, ministries, Sami organisations and conservationists) were called to a meeting to discuss the focus of the synthesis.

The draft protocol and the stakeholder suggestions were then discussed at a meeting of the Review Team and a CEE Systematic Review specialist. During this meeting, the experts on the Review Team confirmed their understanding of the precise scientific scope of the review question and modified the primary question, the choice of search terms and the eligibility criteria appropriately. The SR question finally arrived at was "What are the impacts of reindeer/caribou (*Rangifer tarandus* L.) on mountain and arctic vegetation?"

Following the agreement of the experts on the draft protocol, it was uploaded to the EviEM website to allow for public scrutiny, and stakeholders, in particular, were invited to comment on it. After revision, the final draft protocol was submitted to *Environmental Evidence* for peer review.

4.3 Format for CEE Protocols

The format and template for submitting protocols can be found on the Environmental Evidence website by following the links below

For Systematic Map Protocols

For Systematic Review Protocols

Section 5

Conducting a Search

Last updated: August 11th 2020

Key CEE Standards for Conduct and Reporting

Systematic Reviews and Systematic Maps:

1. **Sources of articles used should capture both commercially published scientific literature and grey literature (may or may not be peer-reviewed).**
2. **Comprehensiveness of the search should be demonstrated by a series of tests using samples of the relevant literature to demonstrate adequate sensitivity.**
3. **All search terms and/or strings, Boolean operators ('AND', 'OR' etc.) and wildcards should be clearly provided (in text or additional files) so that the exact search is replicable by a third party.**
4. **Comprehensive information should be given about the databases and websites searched and search engines used (including any search options or settings chosen), together with dates of searches.**
5. **Any update to searches undertaken during the conduct of the review should be reported and justified.**
6. **A clear account of grey literature and supplementary searches should be provided.**
7. **Limitations due to, for example, language or publication date should be considered.**

5.1 Background

To achieve a rigorous evidence synthesis searches should be transparent and reproducible and minimise biases. A key requirement of a review team engaged in evidence synthesis is to try to gather a maximum of the available relevant documented bibliographic evidence in articles and the studies reported therein. Biases (including those linked to the search itself) should be minimized and/or highlighted as they may affect the outputs of the synthesis (Petticrew & Roberts, 2006; EFSA, 2010; Higgins & Green 2011). Failing to include relevant information in an evidence synthesis could significantly affect and/or bias its findings (Konno & Pullin 2020). This section assumes that full planning has previously been undertaken and the Protocol sets out the plan for the search.

A step-by-step overview of the search process for evidence synthesis is illustrated in Figure 3.1 (Section 3). The planning section (Subsection 3.2) should be read and used in combination with this Section and the overlap is intentional.

5.2 Conducting the Search

Once the search terms and strategy have been reviewed and agreed in the published Protocol, the review team can conduct the search by implementing the whole search strategy.

5.2.1 Prioritizing bibliographic sources

Glanville et al. (in press) suggests that the Review Team should start the search using the source where the largest number of relevant papers are likely to be found, and subsequent searches can

be constructed with the aim to complement these first results. Sources containing abstracts allow greater understanding of relevance and should be given priority. Combined with the use of the test-list, ordering the use of sources may allow the Review Team to find the largest number of relevant articles early during the search, which is useful when time and resources are limited. Searching the grey literature can be conducted in parallel with searches in sources of indexed documents.

5.2.2 Modifying the search string

The list of search terms needs to be combined into search strings that retrieve as many relevant results as possible (exhaustiveness) while also limiting the number of irrelevant results (precision). This will first be done at the Protocol stage (see Section 3). However, search strings needs to be modified (usually simplified) to match the functionality of each electronic bibliographic source to be searched (e.g. Haddaway et al. 2015). To modify the string, the team should consult the **syntax** available in the help pages of the bibliographic sources, including details of the limitations on use of **Booleanoperators**, where applicable. All modification should be fully recorded and reported.

The search syntax is the set of options provided in the interface of the bibliographic source to achieve searches. The syntax options can usually be found in the help pages of the bibliographic source interface.

Typical syntax features are listed below and will vary by interface:

- **Wildcards and truncation**: symbols used within words or at the end of the root of the word to signal that the spelling may vary. Wildcards are useful within words to capture British and US spelling variants, for example 'behavi?r' in some interfaces will retrieve records containing 'behaviour' as well as 'behavior'. As well as wildcards within words, many interfaces offer truncation options at the end of word stems. Truncation can help with identifying words with plural and various grammatical forms. For example, *'forest*'* in some bibliographic sources will retrieve records containing *forest, forests, forestry, forestal…* Some options can also be further defined, for example in the Ovid interface 'forest$1' can be used to restrict searches to words with no or one extra character.
- **Parentheses** are used, where provided, to group search terms together (e.g. a set of synonyms linked by a Boolean operator, see below) and they determine the sequence in which search operations will be carried out by the interface. Search string operations within parentheses are, typically, carried out before those that are not enclosed within parentheses. In complex search strings, nesting of groups of search terms within different sets of parentheses may be helpful, and the search operation is then performed first on the search terms that are within the innermost set of parentheses. In this sense, parentheses as used in search strings function in a similar way to those used in mathematical calculations. For example: (road*OR railway*) AND (killing OR mortality) (for more explanations about OR, see Boolean operators below).

- **Phrase searching:** Some database interfaces allow words to be grouped and searched as phrases by using, for example, double quotation marks. For example, *"organic farming", "tropical forest".*
- **Lemmatization:** lemmatization involves the automated reduction of words to their respective "lemmas" (roots). For example, the lemma for the words "computation" and "computer" is the word "compute". When using defense as a search term, it would also find variants such as defence. Lemmatization can reduce or eliminate the need to use wildcards to retrieve plurals and variant spellings of a word, but it may also retrieve irrelevant variants (e.g. *cite* as a search term may retrieve articles with *citing, cities, cited and citation*, Web of Science helpfile). Web of Science automatically applies lemmatization rules to Topic and Title search queries. This facility is not available in all interfaces.

5.2.3 Refining the results

The finalised search extracts a first pool of articles that is a mixture of relevant and irrelevant articles, because the search, in trying to capture the maximum number of relevant papers, inevitably captures other articles that do not attempt to answer the question. Screening the outputs of the search for eligibility will be done by examining the extracted papers at title, abstract and full-text (See Section 6). If the volume of search results is too large to process within available resources, the Review Team may consider using some tools provided by some electronic databases (e.g. Web of Science) to refine the results of the search by categories (e.g. discipline, research areas) in order to discard some irrelevant articles prior to extracting the final pool of articles and thus lower the number of articles to be screened. There is a real risk in using such tools, as removing articles based on one irrelevant category may remove relevant papers that also belong to another relevant category. This can occur because categories characterise the journal rather than each article and because we are relying on the categories being applied consistently. As a consequence, using refining tools provided by electronic bibliographic sources should be done with great caution and only target categories that are strongly irrelevant for the question (e.g. excluding PHYSICS APPLIED, PERIPHERAL VASCULAR DISEASE or LIMNOLOGY in a search about reintroduction or release of carnivores). Using these tools on the results of a search should not change the number of articles of the test list that have been successfully retrieved. The test-list is again an indicator of the performance of the strategy when using such tools. If the Review Team do decide to use such tools, they should report all details of tools used to refine the outputs of the search prior to screening in the evidence synthesis Protocol and discuss the limitations of the approach they have used.

5.2.4 Searching for grey literature

More and more documents are being indexed including those in the grey literature (Mahood et al. 2014). Nevertheless, conducting a search for grey literature requires time and the authors should assess the need to include it or not in the synthesis (Haddaway and Bayliss 2015). Repeatability and susceptibility to bias should be assessed and reported as much as possible.

Bibliographic tools for grey literature

There are some databases or platforms which reference grey literature. INIST (Institute for Scientific and Technical Information, France) holds the European OpenSIGLE resource (opensigle.inist.fr), which provides access to all the SIGLE records (System for Information on Grey Literature), new data added by EAGLE members (the European Association for Grey Literature Exploitation) and information from Greynet. There are also some programs which can help to make web-based searches for grey literature more transparent, a practice that is part of "scraping methods" (Haddaway 2015). Examples of sources available for grey literature:

- BASE (https://www.base-search.net) allows the selection of document types and provides the option to focus on unpublished material
- eu provides access to more than 700.000 bibliographical references of grey literature produced in Europe.
- Zenodo is an open-access repository initially linked to European projects. It welcomes research outputs from all over the world and all disciplines, including grey literature. It allows search by keywords and includes publications, thesis, datasets, figures, posters, etc.

Examples of sources providing access to theses and dissertations include: DART-Europe (free); Open Access Theses and Dissertations (free); ProQuest Dissertations and Theses (http://pqdtopen.proquest.com/, upon subscription); OAISTER; EThOS (British Library, free); WorldCat.org (free); OpenThesis.org (free, dissertations/theses, but does include other types of publications). Further resources can be found at http://www.ndltd.org/resources/find-etds. Individual universities frequently provide access to their thesis collections.

Websites of organisations and professional networks

Many organisations and professional networks make documents freely available through their web pages, and many more contain lists of projects, datasets and references. The list of organisations to be searched is dependent upon both the subject of the evidence synthesis and any regional focus (see examples in Land et al. 2013; Ojanen et al. 2014; Soderström et al. 2014; Bottrill et al. 2014). Many websites have a search facility but their functionality tends to be quite limited and must be taken into consideration when planning for the time allocated to such task.

Examples:

- TROPENBOS is a non-governmental agency created in the Netherlands in 1986. It contributes to the establishment of research programmes in tropical forestry and it has its own website with many documents, including proceedings of workshops, books and articles that contain useful datasets and references. tropenbos.org
- Databases such as ScienceResearch.com and AcademicInfo.net, contain links to hand-selected sites of relevance for a given topic or subject area and are particularly useful when searching for subject experts or pertinent organisations, helping to focus the searching process and ensure relevance.

Asking authors, experts and colleagues

Direct contact with knowledge-holders and other stakeholders in networks and organisations may be very time-consuming but may allow collection of very relevant articles (Bayliss & Beyer 2015; Schindler et al. 2016). This can be especially useful to help access older or unpublished data sources, when the research area is sensitive to controversy (e.g. GMO, Frampton, pers. comm.) or when resources are limited (Doerr et al. 2015). This may also help enable access to articles written in languages other than English.

World-wide web

Search engines (e.g. Google, Yahoo) cannot index the entire web, and they differ widely in the order of their results. They all have their own algorithms favouring different criteria and both retrieval and ranking of results may be affected by the location, the device used to search (mobile, desktops), the business model of the search engine and commercial purposes. It is important to use more than one search engine to increase chance to identify relevant papers. Google Scholar is often used to scope for existing relevant literature but it cannot be used as a standalone resource for evidence synthesis (see 1.3.2, Bramer et al. 2013; Haddaway et al. 2015)

5.2.5 Additional approaches: hand-searching, snowballing and citation searching

Hand-searching is a traditional (pre-digital) mode of searching which involves looking at all items in a bibliographic source rather than searching the publication using search terms. Hand-searching can involve thoroughly reading the tables of contents of journals, meeting proceedings or books (Glanville, in press).

Snowballing and citation searching (also referred to as ‘pearl growing’, ‘citation chasing’, ‘footnote chasing’, ‘reference scanning’, ‘checking’ or ‘reference harvesting’) refer to methods where the reference lists contained within articles are used to identify other relevant articles (Sayers 2007). Citation searching (or ‘reverse snowballing’) uses known relevant articles to identify later publications which have cited those papers on the assumption that such publications may be relevant for the review.

Using these methods depends on the resources available to the Review Team (access to sources, time). Hand-searching is rarely at the core of the search strategy, but snowballing and citation searching are frequently used (e.g. McKinnon et al. 2016). Recent developments in some bibliographic sources automatically highlight and allow the user to link, to cited and related articles when viewing (e.g. when scanning Elsevier journals, or when downloading full-text PDF). This may be difficult to handle as those references may or may not have been found by the systematic approach using search strings and may have to be reported as additional articles. The use of those methods and their outputs should be reported in detail in the final evidence-synthesis.

5.3 Managing References and Recording the Search

Good documenting, recording and archiving of searches and their resulting articles may save a substantial amount of time and resource by reducing duplication of results and enabling the search to be re-assessed or amended easily (Higgins & Green 2011). Good recording ensures that any of the limitations of the search are explicit and hence allows assessment of any possible consequences of those limitations on the synthesis' findings. Good archiving enables the Review Team to respond to the queries about the search process efficiently. If a Review Team is asked why they did not include an article in their review, for example, proper archiving of the workflow will allow the team to check whether the article was detected by the search, and if it was, why it was discarded.

Good documenting, recording and archiving has two main aspects: (1) the clear recording of the search strategy and the results of all of the searches (records) and (2) the way the search is reported in the evidence synthesis Protocol and final report. Reporting standards keep improving (see a comparative study in Mullins et al. 2014) and many reporting checklists exist to help Review Teams (Rader et al. 2014). See Section 10 for CEE reporting standards.

5.3.1 Keeping track of the search strategy and recording results

The Review Team should document its search methodology in order to be transparent and to be able to justify their use of a search term or the choice of resources. Enough detail should be provided to allow the search to be replicated including the name of the database, the interface, the date of the search and the full search with all the search terms, which should be reported exactly as run (Kugley et al. 2016). The search history and number of articles retrieved by each search should be recorded in a logbook or using screenshots and may be reported in the final evidence synthesis (e.g. as supplementary material). The number of articles retrieved and screened and discarded should be recorded in a flow diagram (see ROSES template) and this should accompany the reporting of the search and eligibility screening stages within an evidence-synthesis report.

For internet searches, reviewers should record and report the URL, the date of the search, the search strategy used (search strings with all options making the search replicable), as well as the number of results of the search, even if this may not be easily reproducible. Saving search results as HTML pages (possibly as screenshots to allow archiving that can be perused later even if the webpage has changed in the meantime) provides transparency for this type of search (Haddaway et al. 2017). Recording searches in citation formats (e.g. RIS files) makes them compatible with reference or review management software and allows archiving for future use.

5.3.2 Reporting the final search strategy and findings

Although the search strategy will have been listed in the Protocol, the searches as finally run should be reported in the final evidence synthesis report, possibly as additional files or supplementary information, since the search as finally run may be different from the Protocol. The final synthesis reports the results and performance of the search. Minor amendments to the

Protocol (e.g. adding or removing search terms) should be reported in the final synthesis, but the search should not be substantially changed once approved by reviewers (but see below).

The Review Team may report the details of each search string and how it was developed (e.g. Bottrill et al. 2014) and whether the strategy has been adjusted to the various databases consulted (e.g. Land et al. 2013, Haddaway et al. 2015) or developed in several languages (e.g. Land et al. 2013). Limitations of the search should be reported as much as possible, including the range of languages, types of documents, time-period covered by the search, date of the search (e.g. Land et al. 2013; Söderström et al. 2014), and any unexpected difficulty that impacted the search compared to what was described in the Protocol (e.g. end of access, Haddaway et al. 2015).

5.4 Updating and Amending Searches

Updating or amending a search may be conducted by the same Review Team that undertook the initial searches, but this is not always the case. Therefore, it is important that the original searches are well documented and, if possible, libraries (e.g. EndNote databases) of retrieved articles are saved (and, if possible, reported or made available) to ensure that new search results can be differentiated from previous ones, as easily as possible.

There are two main reasons why a search needs to be changed. The first may occur when the evidence synthesis extends over a long time period (for instance more than 2 years) and the publication rate of relevant documents on the topic is high. In this case, the conclusions of the review may be out of date even before it is published. It is recommended that the search is rerun using the same search strategy (Bayliss et al. 2016) for the time period elapsed subsequent to the end of the initial search and before the report is finalised. The second case occurs when the evidence synthesis final report has already been published, and there is a need for revision because new primary research data or developments have subsequently been published and need to be taken into account. In this case the search Protocol should be checked to identify whether new search terms need to be added or additional sources need to be searched. Deciding whether a new Protocol needs to be published will depend on the extent of the amendments and may be discussed with the Collaboration for Environmental Evidence. From the moment a search is completed, new articles may be published as research effort is dynamic.

There are a number of issues that need to be considered when updating a search:

- Do you have access to the original search strings, sources, and can you read these files (proper software available)?
- Was the original search Protocol adequate and appropriate or does it need revising?
- Do you know when the initial search took place and which time boundaries were set up at that time? If not, can you contact the authors to get those details?
- If relevant, do you have similar details regarding searches in grey literature?
- Do you have access to the same sources of documents (e.g. database platforms), including institutional websites, subscriptions?
- Will the same languages be used?

Then the revised (or original) strategy may be run (Bayliss et al. 2016). As with the original searches, it is important to document clearly any updates to the searches, their dates, and any reasons for changes to the original searches, most typically in an appendix. If the new search differs from the initial one, a new Protocol may need to be submitted before the amendment is conducted (Bayliss et al. 2016).

Section 6

Eligibility screening

Last updated: August 11th 2020

CEE Standards for conduct and reporting

1. **Eligibility criteria should be precisely defined (e.g. reliance on broad and potentially ambiguous terms should be avoided) and all key elements of the question considered.**
2. **Eligibility criteria applicable to all stages of screening should be provided.**
3. **Eligibility criteria should be consistent between a-priori Protocol and review or differences fully explained.**
4. **Eligibility criteria should be independently applied by more than one reviewer, ideally to all articles screened at title, abstract and full text stages. Pragmatic decisions about dual screening of subsamples only may be acceptable when large numbers of articles are screened. In such a case the rationale and methods of subsampling should be fully described and justified (see section 6.3.3)**
5. **Consistency of screening decisions should be measured and reported and all disagreements between reviewers discussed and resolved so as to inform subsequent decisions.**
6. **The number of unique articles found during the search (after removal of duplicates) should be reported and the number excluded at each stage of the screening process fully presented (e.g. in a flow diagram or table).**
7. **The reasons for exclusion of each article/study considered at full-text should be reported (e.g. in additional files).**
8. **A list should be provided of any articles which had unclear eligibility status after completion of full-text screening (with explanation why they could not be classified) and of any articles that could not be obtained for full-text screening.**
9. **The final list of studies eligible for (included in) the review should be provided.**

6.1 Background

The eligibility screening step of a systematic review or systematic map (which may also be referred to as 'study selection', 'evidence selection' or 'inclusion screening') involves the application of *eligibility criteria* that determine which of the primary research studies identified in searches are relevant for answering the review or map question; and the use of a

systematic *screening process* for applying the eligibility criteria to the search results in such a way as to minimise the risk of introducing selection bias (McDonagh et al., 2013). Both the eligibility criteria and the screening process should be planned in advance (Section 3) and specified in the evidence synthesis Protocol (Section 4).

6.2 Removing duplicates

As a first step in screening, duplicate articles, that are common in search results, should be removed where possible before eligibility screening starts. Inclusion of duplicates in an evidence synthesis could lead to double-counting of data, which might introduce bias (Tramèr et al., 1997), as well as creating unnecessary additional screening effort. Many reference management tools enable automated identification and removal of duplicate articles (e.g. 'fuzzy matching' of references in Eppi Reviewer and this may be particularly helpful if large numbers of duplicates are present. However, care should be taken to avoid inadvertently removing articles which are not duplicates. If an automated process is used for identifying duplicates it should not be assumed that this will always classify the articles accurately.

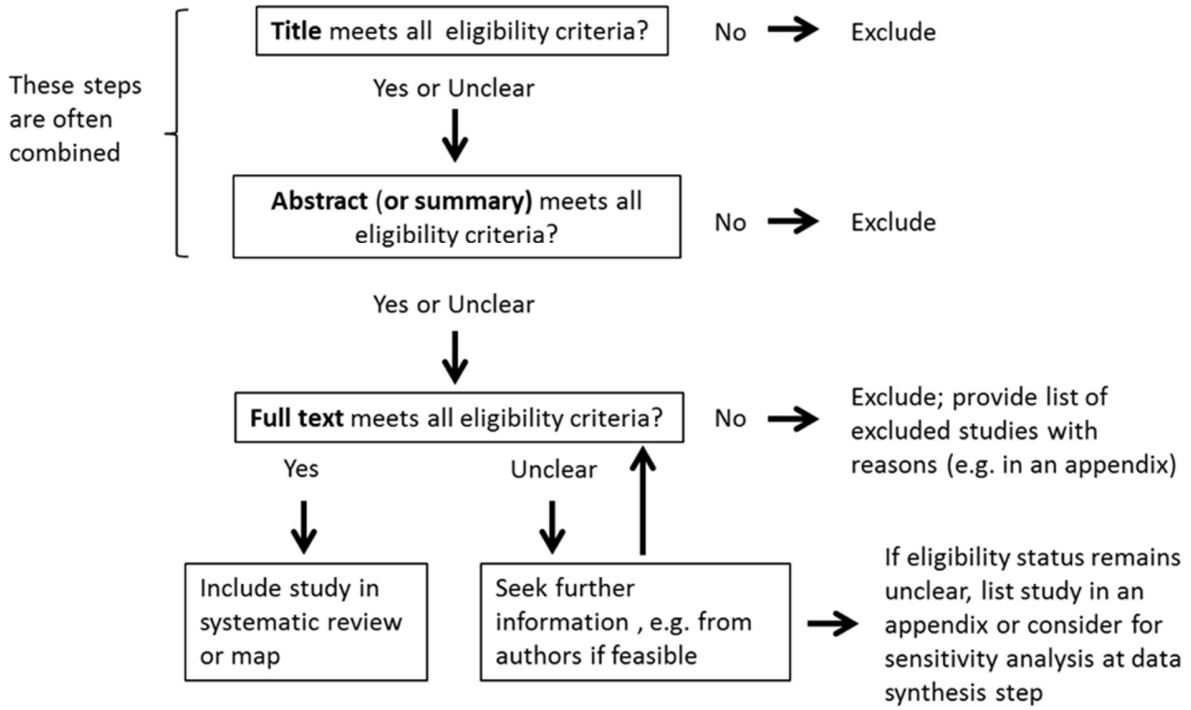
6.3 The screening process

6.3.1 Rationale and overview of the screening process

The process of eligibility screening aims to ensure that the eligibility criteria are applied consistently and impartially so as to reduce the risk of introducing errors or bias in an evidence synthesis. Articles identified in searches are typically structured in having a title, an abstract (or summary), and/or a 'full text' version such as an academic journal paper, agency report, or internet pages. Eligibility screening can be applied at these different levels of reading to impose a number of filters of increasing rigor and thus screening is normally a stepwise process. CEE recommends that at least two filters are applied: (i) a first reading of titles and abstracts to efficiently remove articles which are clearly irrelevant; and (ii) assessment of the full-text version of the article.

Depending on the nature of the evidence synthesis question and the number of articles requiring screening, titles and abstracts may be screened separately or together. If only an insignificant number of articles can be excluded on title alone (e.g. as found in a systematic review of the environmental impacts of poverty rights regimes by Ojanen et al. 2017), then combining the title and abstract screening in a single step may be more efficient. In cases where insufficient information is available in the title or abstract to enable an eligibility decision to be made, or if the abstract is missing, then the full-text version should be obtained and examined. An overview of the eligibility screening process is shown in Figure 6.1.

Figure 6.1. The eligibility screening process for systematic reviews or systematic maps



As shown in Figure 6.1, the screening process starts with individual articles but final eligibility decisions are made at the level of studies, taking into account any linked articles that refer to the same study (see '*Identifying linked articles*' below). The evidence selection decision process is conservative at each step so that only articles which do not meet the inclusion criteria are excluded; in any cases of doubt, articles proceed to the next step for further scrutiny. If after full-text screening the eligibility of a study remains unclear, further information should be sought, if feasible (e.g. by contacting the authors), to enable the study to be included or excluded. Any studies whose eligibility still remains unclear after this process should be listed in an appendix to the systematic review or systematic map report. In systematic reviews, an option could be to include studies of unclear relevance in a sensitivity analysis. The approach for handling unclear studies should be considered during Protocol development and specified in the systematic review or systematic map Protocol.

A single set of eligibility criteria can be used to screen titles, abstracts and full-text articles (e.g. Rodriguez et al. (2016) used the eligibility criteria shown in Box 3.2 for screening titles and abstracts and then applied the same criteria to full-text articles). However, if the information reported in titles and abstracts is limited it may be efficient to use a smaller subset of the eligibility criteria to screen the titles and/or abstracts, and apply the more detailed full set of eligibility criteria for the screening of full-text articles. Whichever approach is used, the eligibility criteria applied at each step should be clearly stated in the Protocol.

6.3.2 Identifying linked articles

If the same data are included more than once in an evidence synthesis this can introduce bias (Tramèr 1997; von Elm et al., 2003; Choi et al., 2014). Therefore, the unit of analysis of interest

in a systematic review or map is usually individual primary research studies (e.g. observational studies, surveys, or experiments), rather than individual articles.

Investigators often report the same study in more than one article (e.g. the same study could be reported in different formats such as conference abstracts, reports or journal papers, or in several different journal papers; von Elm et al. 2004), and we refer to these as 'linked articles'. Although there is often a single article for each study, it should never be assumed that this is the case (Higgins & Green, 2011). Linked articles may range from being duplicates (i.e. they fully overlap and do not contribute any new information) to having very little overlap. Articles which are true duplicates should be removed to avoid double-counting of data. The remaining linked articles which refer to a study should be grouped together and screened for eligibility as a single unit so that all available data pertinent to the study can be considered when making eligibility decisions.

It may be difficult to determine whether articles are linked, as related articles do not always cite each other (Bailey 2002; Barden et al., 2003) or share common authors (Gøtzsche 1989). Some 'detective' work (e.g. checking whether the same data appear in more than one article, or contacting authors) may therefore be needed by the review team. Although it would be ideal to identify linked articles that refer to the same study early on the screening process, it may only become clear at the full-text screening stage that articles are linked. Once the links between articles and studies have been identified, a clear record will need to be kept of all articles which relate to each study. This may be done using a separate document or spreadsheet, or using grouping or cross-referencing functions available in bibliographic reference management tools.

6.3.3 Number and expertise of screeners

Eligibility decisions involve judgement and it is possible that errors or bias could be introduced during eligibility screening if the process is not conducted carefully.

Possible problems that could arise at the eligibility screening step are:

- Some articles might be misclassified due to the way members of the review team interpret the information given in them in relation to the eligibility criteria;
- One or more articles might be missed altogether, due to human error;
- Review team members may (knowingly or not) introduce bias into the selection process, since human beings are susceptible to *implicit bias* and experts in a particular topic often have pre-formed opinions about the relevance and validity of articles (e.g. Higgins & Green, 2011; Gøtzsche & Ioannidis, 2012).

Appropriate allocation of the review team to the eligibility screening task, in terms of the number and expertise of those involved, is important to ensure efficiency (DEFRA, 2015) and can help to minimise the risk of errors or bias. If any members of the review team are authors of articles identified in the searches then the allocation of screening tasks should ensure that members of the review team do not influence decisions regarding the eligibility of their own articles.

Number of screeners

It has been estimated that when eligibility screening is done by one person, on average 8% of eligible studies would be missed, whereas no studies would be missed when eligibility screening is done by two people working independently (Edwards et al., 2002). To ensure reliability of the eligibility screening process, articles providing guidance on conducting systematic reviews in environmental research (EFSA, 2010; Rooney et al., 2014; Sargent & O'Connor, 2014) and health research (CRD, 2009; Higgins & Deeks, 2011; McDonagh et al., 2013) recommend that eligibility screening should be performed where possible by at least two people. The screeners need not necessarily be the same two people for all articles or for all screening steps. Options could be for one person to screen the articles and the second person to then check the first screener's decisions; or both screeners may independently perform the selection process and then compare their decisions. Independent screening is preferable since it avoids the possibility that the second screener could be influenced by the first screener's decision.

A potential problem with eligibility screening being conducted by a single screener is that any errors in the classification of articles by the screener, or any articles missed from classification, may go undetected, if checking by a second screener is not done on an adequate number of articles. Reliability checking can be done (e.g. using screener agreement statistics) but this has limitations which should be taken into consideration, as we explain below see '*Assessing screener agreement*').

Eligibility screening can be a time-consuming process, typically taking an hour or more for a screener to assess 200 titles or 20 abstracts (DEFRA, 2015). If the evidence base is extensive such that large numbers (e.g. tens of thousands) of articles would need to be screened, it might not always be feasible for two or more screeners to work on all screening steps. Consideration may then need to be given as to whether the systematic review or systematic map question, or the eligibility criteria, should be refined (e.g. narrowing the scope) to make the evidence synthesis manageable within the available resources (see Section 2). Discussion with relevant stakeholders, e.g. research commissioners, may be helpful in resolving any difficulties if the level of rigor expected of eligibility screening will be difficult to achieve within the available resources. Employing a single screener at one or more steps of the eligibility screening process, subject to checking screener reliability, is a pragmatic approach which may be justifiable on a case-by-case basis depending on the nature of the topic and how critical it is to minimise the risk of selection bias (e.g. Langer et al., 2017), but should not be considered as being reflective of best practice (see '*Assessing screener agreement*' below). It may be tempting to consider employing a single screener for titles, since the information available in a title is usually relatively limited and titles can often indicate that an article is irrelevant without the need to expend detailed effort in screening (DEFRA, 2015). However, selection bias could arise at title screening (just as it could at abstract or full-text screening) if a screener is not impartial, and this could be especially important for evidence syntheses on contentious topics. Furthermore, in our experience it is not uncommon for a small proportion (~1%) of articles to be completely missed from screening by a single reviewer, due to human error (e.g. screener fatigue when assessing thousands of articles). For these reasons, good practice would be to employ a minimum of two screeners at the title screening as well as abstract and full-text screening steps.

For systematic maps the need to minimise selection bias may seem less critical than for systematic reviews, since the output and conclusions of systematic maps are often descriptive. Nevertheless, an underlying expectation of systematic maps is that the searching and eligibility screening steps should be conducted with the same rigor as for systematic reviews (James et al., 2016). It is therefore good practice in all types of evidence synthesis that at least two people conduct eligibility screening of each article. We recommend that deviations from this should only be made as exceptions, where clear justification can be provided, and is agreed among all relevant stakeholders. This is important for maintaining the integrity of systematic evidence synthesis as a 'gold standard' or 'benchmark' approach for minimising the risk of introducing errors or bias, and to avoid creating confusion as to whether the methods employed in specific evidence syntheses truly constitute those of a systematic review or systematic map, rather than, for example, a traditional literature review or rapid evidence assessment (DEFRA, 2015).

If a pragmatic decision is made by the review team to proceed with a systematic review or systematic map involving a large number of articles to screen and to use only one screener for some of the articles then, for consistency with good practice as defined above, the following information should be provided in the Protocol and final evidence synthesis report:

- a clear justification for using one screener to screen all and a second to screen only a sample, stating which steps of the screening process this will be applied to;
- evidence of the reliability of the approach (i.e. the reliability of the screener's decisions should be tested and reported; see '*Assessing screener agreement'* below);
- acknowledgement that the use of one screener to screen all and a second to screen only a sample at one or more steps of eligibility screening is a limitation (this should be stated in the conclusions section, critical reflection or limitations section, and, if possible, also in the abstract).

Ultimately, it is the review team's responsibility to ensure that, where possible, methods are used which minimise risks of introducing errors and bias, and that any limitations are justified and transparently reported.

Expertise of screeners

There is no firm 'rule' about how many of the screeners should be topic experts. Given the complexity of environmental topics it is important that the team has adequate expertise in evidence synthesis and the question topic to ensure that important factors relating to the evidence synthesis question are not missed (DEFRA, 2015). However, topic experts may lack impartiality as they are likely to be very familiar with the literature relevant to the evidence synthesis question which may risk selective screening decisions being made (Gøtzsche and Ioannidis 2012). A pragmatic approach to reduce the risks of any conflicts of interest within a review team could be to include screeners with different backgrounds and expertise, to ensure diversity of stakeholder perspectives.

6.3.4 Assessing screener agreement

An assessment of agreement between screeners during pilot-testing can help to ensure that the eligibility screening process is reproducible and reliable (Frampton et al, 2017). If necessary, the eligibility criteria and/or screening process may be modified and re-tested to improve the agreement between screeners as long as deviation from the Protocol is explained and justified. Agreement can be assessed by: recording the observed proportions of articles where pairs of screeners agree or disagree on their eligibility decisions; calculating a reviewer agreement statistic; and/or descriptively tabulating and discussing any disagreements.

A widely used statistic for assessing screener agreement is Cohen's kappa (Altman, 1991), which takes into account the level of agreement between screeners that would occur by chance. But interpretation of kappa scores is subjective since there is no consensus as to which scores indicate 'adequate' agreement, and the concept of 'adequate' agreement is itself subjective.

To assess screener agreement, a sample (as large as possible) of the articles identified in searches should be screened by at least two people and their agreement determined. The size of the sample should be justified by the review team and the articles comprising the subset should be selected randomly to avoid bias towards certain authors, topics, years or other factors.

Use of a kappa statistic to guide pilot-testing of eligibility screening where two or more people will screen each article is a pragmatic approach to optimise efficiency of the process, in which case the limitations of the agreement statistic and its somewhat arbitrary interpretation are not critical. However, use of the kappa statistic to demonstrate high reviewer agreement in support of employing only one screener to assess the majority of articles is not advised. The potential insensitivity of overall screener agreement measures to specific discrepancies in screener agreement suggests that a kappa statistic might not be adequate as a justification that a single screener has sufficient reliability in their screening decisions to protect against the risk of introducing errors or selection bias.

As there is no consensus on what 'adequate' rates of agreement are (unless reaching 100%), the review team should justify the level of agreement reached and explain in the evidence synthesis report whether relying on a single screener may have led to any relevant studies being excluded. If so, an explanation should be given as to how this would affect interpretation of the evidence synthesis conclusions. Presentation of a decision matrix showing the combinations of screener agreements may be helpful to support any discussion and interpretation of screener reliability.

6.3.5 Resolving disagreements

A process for resolving any disagreements between screeners should be agreed by the review team and to ensure consistency this should be pre-specified in the Protocol. An approach which appears to be commonly used (Peterson and bin Ali, 2011), and which works efficiently in our experience, is that the screeners meet to discuss their disagreements to reach a consensus; if consensus is not reached a third opinion could then be sought, from another member of the review team or the project advisory group. The exact approach is a matter of preference; for example, abstracts over which there is disagreement could be discussed by the screeners before

proceeding to the full-text screening step (to avoid obtaining full-text articles unnecessarily), or the articles could be directly passed to the full-text screening step (to enable decisions to be based on all available information). Records of all screening decisions should be kept to ensure that, if necessary, the review team can justify their study selection. Screening decisions can often be recorded conveniently in user-definable fields in reference management tools. Pilot-testing the screening process, described below, can be helpful to identify whether some screeners differ systematically from others in the eligibility decisions they make.

6.4 Recording and documenting eligibility screening

6.4.1 Reporting the eligibility criteria, screening process and screening results

A concise summary of the eligibility criteria and screening process should be given in the final report, including in the abstract or summary. An explanation must be provided in the final report if the methods employed differed from those specified in the Protocol. It is particularly important to consider whether any changes to the Protocol could have introduced errors or bias.

It is good practice to include a flow diagram in the evidence synthesis report to show how many unique articles were identified (i.e. after removing any duplicates), and to indicate how many of these were excluded at the title, abstract and full-text screening steps (CRD, 2009; Liberati et al., 2009; EFSA, 2010; Higgins & Green, 2011; McDonagh et al., 2013; Rooney et al., 2014). The flow diagram should also clarify the relationship between articles and studies so that it is clear how many articles and unique studies were included in the systematic review or systematic map; and should give reasons why any studies were excluded at the full-text selection step. A template flow diagram is provided by the ROSES reporting standards.

The flow diagram template may be adjusted to display how the eligibility screening was conducted. For example, the diagram may be expanded to accommodate further panels if titles and abstracts are screened separately. In addition to the flow diagram, a list of the studies which were excluded at the full-text screening step should be provided, indicating the reasons for exclusion (e.g. as an appendix to the evidence synthesis report). Whilst the ROSES template indicates the minimum information on the results of eligibility screening that should be reported, some authors advocate specifying further information. For example, the flow diagram could include an indication of how many of the included studies contributed to any meta-analyses (e.g. Sagoo et al., 2009), or an indication of how many studies informed quantitative and qualitative analyses for the primary outcome of interest (e.g. Liberati et al., 2009; Higgins & Green, 2011).

Any definitions and instructions on interpretation of the eligibility criteria used by the review team should be reported at least in the Protocol. Details of the screening process which should be documented in the Protocol and also stated concisely in the evidence synthesis report are: the number of screeners involved at each eligibility screening step; whether screening decisions were independent; the expertise of the screeners; the pilot-testing process; any assessments of screener agreement, with justification for the methods chosen; the process employed for resolving any screener disagreements; and how any missing or unclear information was handled.

Any limitations in the eligibility screening process should be mentioned in the Discussion (or Critical Reflection) section of the final evidence synthesis report so that readers can consider

them when interpreting the overall findings of the evidence synthesis (Liberati et al., 2009). If there are any serious limitations in the eligibility screening criteria or the screening process which could affect the overall conclusions of the systematic review or systematic map these should, where possible, also be mentioned in the abstract or summary.

6.4.2 Keeping an archive of screening decisions

It is important that a record is kept of all eligibility screening decisions so that judgements made during conduct of the systematic review or systematic map are transparent and, if necessary, defensible (e.g. if any readers query why a particular study was not included). A record of the screening decisions should be saved (e.g. in a reference management tool or relational database) that can easily be interrogated to display articles which were included, excluded, or deemed unclear at each selection step. The tool or database containing the full set of screening decisions should be archived in such a way that it can be made available if requested by any readers of the systematic review or systematic map report.

Section 7

Data Coding and Data Extraction

Last updated: August 11th 2020

CEE Standards for conduct and reporting

Systematic Reviews:

1. **Methods by which raw data from each study were coded and extracted should be stated in the Protocol so that the process can be replicated and confirmed in the final report unless deviations are reported and justified.**
2. **All data coded or selected for extraction should be provided in a table or spreadsheet as set out in the *a-priori*Protocol (this includes data used in the synthesis for each study, e.g. outcome metrics or effect size, and meta-data).**
3. **Data coded or extracted from each study should be cross checked by at least two independent reviewers. If not, an explanation should be provided of how a sample of coded or extracted data was cross checked between two or more reviewers.**
4. **Any process for obtaining and confirming missing or unclear information or data from authors should be described.**

Systematic Maps:

1. **Methods by which raw data from each study were coded should be stated in the Protocol so that the process can be replicated and confirmed in the final report unless deviations are reported and justified.**
2. **All data coded should be provided in a table or spreadsheet as set out in the *a-priori* Protocol.**
3. **Data coded from each study should be cross checked by at least two independent reviewers. If not, an explanation should be provided of how a sample of coded data was cross checked between two or more reviewers.**
4. **Any process for obtaining and confirming missing or unclear information or data from authors should be described.**

Systematic reviews and systematic maps are based on data that are extracted systematically and transparently from each eligible study using procedures that are sufficiently well documented to allow other reviewers to obtain the same data from the same studies. The term 'data' is used here to mean any information about (or deriving from) a study, including details of methods, location or setting, context, interventions, outcomes, and results (Higgins and Green 2011).

Data coding and data extraction refer to the process of systematically extracting relevant information from the articles included in the Evidence Synthesis. Data coding is the recording of relevant characteristics (meta-data) of the study such as when and where the study was conducted and by whom, as well as aspects of the study design and conduct. Data coding is undertaken in both Systematic Reviews and Systematic Maps. Data extraction refers to the recording of the results of the study (e.g. in terms of effect size means and variances or other important findings). Data extraction is undertaken in Systematic Reviews only (see also Section 3.5). The precise order in which data coding, critical appraisal and data extraction are undertaken varies from one Systematic Review to another. In our experience, there is frequently an iterative relationship between them and they are often conducted together. Therefore our advice is to read through both this Section and Section 8 before proceeding.

Coded and extracted data should be recorded on carefully designed forms and undertaken with the appropriate synthesis in mind (see Section 9). Great care should be taken to standardise and document the processes of data coding and data extraction, the details of which should be reported to increase the transparency of the process. Because each review is different, data collection forms will vary across reviews. However, there are many similarities in the types of information that are important, and forms can be adapted from one review to the next. To some extent data coding and data extraction should be guided by *a priori* rules described in the Protocol, but the complexity of the operation means a degree of flexibility may be maintained. Sensitivity analyses can be used to investigate the impact of coding and extracting data in different ways when there is doubt about the optimum method.

7.1 Assessing agreement between data coders/extractors

An assessment of agreement between members of the review team tasked with data extraction during pilot-testing can help to ensure that the process is reproducible and reliable as it is for screening (Frampton et al 2017). Ideally, data extraction should be piloted on a sample of relevant studies at the planning stage (see Section 3). However, data extraction outlined in the Protocol may need to be modified following assessment and re-tested to improve the agreement between team members.

It is difficult to perform formal statistics on the repeatability of data extraction, but some attempt to verify repeatability should be made. A second reviewer should at least check a random subset of the included studies to ensure that the *a priori* rules have been applied or the rationale of deviations explained. Randomly checking team members' interpretation of data extraction in the Protocol acts as a check on data hygiene and human error (e.g. misinterpretation of a standard error as a standard deviation). Where data extraction has limited repeatability it is desirable to maintain a record of exactly how the extraction was undertaken on a study by study basis. This maintains transparency and allows authors and other interested parties to examine the decisions made during the extraction process. Particular attention should be paid to the data used to generate effect sizes. For transparency, data extraction forms should be included in an appendix or supplementary material.

7.2 Data coding

Provided sufficient planning has been undertaken at the Protocol stage (See Section 3.5), data coding should be a relatively straightforward task involving careful reading of the full text of each study. Variables or characteristics to be coded for each study should be included in a suitable spreadsheet prior to coding. Although the list of coded variables should have been discussed with stakeholders at the planning stage, there will usually be a need to refine definitions and discuss details of how each variable should be coded once the studies are read at full text. Decisions taken at this stage should be fully reported.

For Systematic Reviews, some of the variables coded will be potential effect modifiers that may cause heterogeneity in effect and therefore need special consideration in terms of how they are recorded so as to facilitate their inclusion in further analysis such as subgroup analysis and meta-regression (see Section 9). Some variables may be categorical whilst others will be continuous. In some cases, quantitative variables may need to be recorded as means and variances in the same way as effect sizes.

For Systematic maps, some of the variables may be used to sort studies into subgroups for data visualisation. Potential methods of data visualisation should be fully considered in advance of data coding so that the necessary information is recorded. Table 7.1 shows an example of a coding sheet from a systematic map on human health impacts resulting from exposure to alien species in Europe (Bayliss et al 2017).

Table 7.1 Example of a coding sheet for a systematic map on human health impacts resulting from exposure to alien species in Europe (Bayliss et al 2017)

Category	Type of data
1. Bibliographic information	a. Publication type b. Year
2. Information relating to the inclusion criteria	a. Population i: Human population affected b. Population ii: Location of exposure c. Population iii: Location of reported impact d. Population iv: Activity of population at exposure e. Exposure i: Taxonomic group of the alien species f. Exposure ii: Species name (binomial) of the alien species g. Exposure iii: Biome at location of exposure h. Exposure iv: Habitat at location of exposure i. Outcome i: Type of human health impact (disease or pathogen transmission/allergen or irritant) j. Outcome ii: Specific condition (type of injury, allergy, dermatitis, disease) k. Outcome iii: Change in human health impact (occurrence/frequency/severity)
3. Information relating to the study	a. Study type (e.g. patient case study, RCT) b. Study design (sampling size, etc.) c. Spatial scale of reported impact d. Comparators e. Timescale f. Other factors affecting the outcome
4. Additional information relating to the species of concern	a. Geographic origin (native range) b. Pathway(s) of introduction (following the classification of Hulme et al. [45])

7.3 Data extraction

When adapting or designing a data extraction form, review authors should first consider how much information should be collected. Extracting too much information can lead to forms that are longer than original study reports, and can be very wasteful of time. Extraction of too little information, or omission of key data, can lead to the need to return to study reports later in the review process.

Sensitivity analyses can be used to investigate the impact of extracting data in different ways when there is doubt about the optimum extraction method. When extracting data from

quantitative studies, it is standard practice to extract the raw or summary data from included studies wherever possible, so a common statistic can be calculated for each study. The results of studies included in a review may take different numerical or statistical forms, which may involve transforming results into a common numerical or statistical measure if possible. In a review of effectiveness which incorporates meta-analysis these results would be pooled to provide a single estimate of effect size (see Section 9). It is important to extract data that reflect points of difference and any heterogeneous characteristics between studies that might affect data synthesis and interpretation of the findings. Whether statistical data synthesis can be performed will depend largely on the heterogeneity of the variables of interest across included studies.

Good practice for data extraction could involve the following steps, which improve transparency, repeatability and objectivity:

- Report the location of study data within each article and means of extraction if data are located within figures.
- Provide the pre-tested data extraction form.
- Data extraction by multiple reviewers using a subset of eligible studies and checking for human error/consistency.
- Include appendices of extracted information.
- Detail contact made with authors requesting study data where they are missing from relevant articles.
- Describe any pre-analysis calculations or data transformations (e.g. standard deviation calculation from standard error and sample size (e.g. Felton et al. 2010 and Smith et al. 2010), and calculation of effect sizes.

Table 7.2 shows and example of a data extraction table from a systematic review on the influence of a reduction of planktivorous and benthivorous fish on water quality in temperate eutrophic lakes (Bernes et al 2015)

Table 7.2 Data extraction table from Bernes et al (2015)

Comparison of responsive and unresponsive lakes

	Responsive lakes (significant improvement)			Unresponsive lakes (no significant improvement)		
	Mean	95% C.I.	n	Mean	95% C.I.	n
Response: Secchi depth 1–3 years after manipulation (vs. before manipulation)						
Lake area (ha)	18	10 – 32	19	40	14 – 119	8
Mean depth (m)	1.7	1.3 – 2.2	19	1.8	1.3 – 2.5	8
Retention time (days)	171	63 – 461	11	409	221 – 754	5
Pre-manipulation TP (µg/l)	144	93 – 223	19	127	62 – 260	6
Mean atmospheric temperature (°C)	8.0	7.0 – 9.0	19	7.8	6.4 – 9.2	8
Duration of main manipulation (yr)	2.1	1.6 – 2.5	19	2.4	1.5 – 3.3	8
Fish removal (kg/ha)	233	160 – 338	18	251	198 – 317	7
Fish removal (kg/ha/yr)	124	84 – 183	18	119	70 – 203	7
Fish stock depletion (%)	56	37 – 76	13	40	17 – 62	5
Response: Chlorophyll *a* 1–3 years after manipulation (vs. before manipulation)						
Lake area (ha)	12	5 – 36	12	35	14 – 86	15
Mean depth (m)	1.3	1.1 – 1.6	11	1.8	1.4 – 2.4	15
Retention time (days)	78	22 – 275	4	210	103 – 428	9
Pre-manipulation TP (µg/l)	196	120 – 322	12	126	83 – 190	13
Mean atmospheric temperature (°C)	8.1	7.1 – 9.1	12	8.2	7.5 – 8.9	15
Duration of main manipulation (yr)	2.0	1.4 – 2.6	12	2.0	1.5 – 2.5	15
Fish removal (kg/ha)	250	149 – 420	11	272	184 – 403	12
Fish removal (kg/ha/yr)	137	74 – 252	11	140	98 – 202	12
Fish stock depletion (%)	78	56 – 99	7	58	42 – 75	12

Data are based on the selected dataset, with stocking-only interventions excluded. Lake areas, mean depths, retention times, pre-manipulation TP concentrations and fish removals were log-transformed before calculation of means and confidence intervals, and then back-transformed.

At this stage, it may be necessary to exclude studies that are seemingly relevant but do not present data in extractable format (e.g. if they do not report standard deviations for control and treatment group(s) or the information required to calculate the statistic). If possible, authors of such studies should be contacted and asked whether they can provide data in a suitable format. Contacting authors for data is not normal practice in environmental science and can be met with surprise and indignation, but it is important to develop the culture and expectation of data accessibility, particularly when the research was publicly funded.

In some cases, where the information required is not presented and cannot be obtained from authors, data can be converted into an appropriate form without problems. For example, it is relatively straightforward to substitute standard deviation for standard errors, confidence intervals, *t*-values, or a one-way *F*-ratio based on two groups (Lipsey & Wilson 2001, Deeks et al. 2005). Where missing data cannot be substituted, it can be imputed by various methods. Imputation is a generic term for filling in missing data with plausible values. These are commonly derived from average or standardised values (Deeks et al. 2005), but also from bootstrapped confidence limits (Gurevitch & Hedges 2001) or predicted values from regression models (Schafer 1997). Alternatively, data points can be deleted from some analyses, particularly where covariates of interest are missing. Such pragmatic imputation or case deletion should be accompanied by sensitivity analyses to assess its impact.

The impacts of imputation or case deletion can be serious when they comprise a high proportion of studies in an analysis. Case deletion can result in the discarding of large quantities of information and can introduce bias where incomplete data differ systematically from complete (Schafer 1997). Likewise, imputing average values or predicted values from regressions distorts covariance structure resulting in misleading p-values, standard errors and other measures of uncertainty (Schafer 1997). Where more than 10% of a data set is missing serious consideration should be given to these problems. More complex imputation techniques are available (see Schafer 1997) and should be employed in consultation with statisticians. If this is not possible, the results should be interpreted with great caution and only presented alongside the sensitivity analysis. Section 9 discusses data analysis in greater detail.

Section 8

Critical appraisal of study validity (Systematic Reviews)

Last updated: August 11th 2020

CEE Standards for conduct and reporting

1. **An effort should be made to identify all relevant sources of bias (threats to internal and external validity)**
2. **Each relevant type of bias (threat to internal and external validity) should be assessed individually for all included studies**
3. **Results should be reported using a critical appraisal sheet constructed and tested at the protocol stage.**
4. **Critical appraisal criteria should be consistent between a-priori Protocol and review or differences fully explained.**
5. **At least two people should have independently critically appraised each study with disagreements and process of resolution reported.**
6. **A description should be provided of how the information from critical appraisal was used in synthesis.**

8.1 Background

Some primary studies provide evidence of higher reliability and relevance than others in respect to the review question. Assessing the comparative validity of the included studies (often referred to as critical appraisal) is of key importance to the resulting value of the Systematic Review (see examples in Box 8.1 and Table 8.1). It can form a basis for the differential weighting of studies in later synthesis or partitioning of studies into subgroups for separate analyses.

Study validity assessment requires a number of decisions about the absolute and relative importance of different sources of bias and data validity elements common to environmental data, particularly the appropriateness of temporal and spatial scales. It is therefore vital that the assessment process be standardised and as transparent and repeatable as possible. This challenge has been extensively covered in the Planning Section (Section 3). Some extra points are made below that may help in the conduct as well as the planning stages.

8.2 Internal validity

In an ideal world, each data set included in a SR should be of high internal validity, thus ensuring that the potential for error and bias is minimised and that any differences in the outcome measure between experimental groups can be attributed to the exposure or intervention of interest. To determine the level of confidence that may be placed in selected data sets, the methodology employed to generate each one must be critically appraised, using a transparent and consistent framework, to assess the extent to which it is likely to prevent systematic errors or bias (Moher et al. 1995). However, the nature of the critical appraisal and the hierarchy employed is dependent on the nature of the question and the 'theory of change'. The Review Team should be able to justify their approach and not blindly follow an established methodology.

In the health sciences, a hierarchy of research methodology is recognised that scores the value of the data in terms of the scientific rigour; the extent to which the methodology seeks to minimise error and bias (Stevens & Milne 1997). The hierarchy of methodological design can be viewed as generic and has been translated from medicine to environmental sciences (Pullin & Knight 2003), but these generic hierarchies are crude tools and usually just a starting point and can rarely be used without modification to ensure relevance to individual review questions. Where a number of well-designed, high-validity studies are available, others with inferior methodology may be demoted from subsequent quantitative analysis to narrative tabulation, or rejected from the SR entirely. However, there are dangers in the rigid application of hierarchies as the importance of various methodological dimensions within studies will vary, depending on the study system to which an intervention is being applied. For example, a rigorous methodology, such as a randomised controlled trial (RCT), applied over inadequately short time and small spatial scales could be viewed as superior to a time series experiment providing data over longer time and larger spatial scales that were more appropriate to the question. The former has high internal validity but low external validity or generalisability in comparison to the latter. This problem carries with it the threat of misinterpretation of evidence. Potential pitfalls of this kind need to be considered at this stage and explored in covariate analyses (e.g. experimental duration

or study area: see Downing et al. 1999 and Côté et al. 2001, respectively) or by judicious use of sensitivity analysis at the synthesis stage (see below).

As a consequence, authors may use existing checklists of the *CEE Critical Appraisal Tool* or other tools as a basis for their specific exercise, but they should either explain why they use them as such (no modification, because not considered to be needed, and why) or adapt them to their own case-study review, in which case the decisions made must be stated and justified (see Gough et al. 2012).

We suggest that review-specific *a priori* assessment criteria for appraising the internal validity are included in the Protocol and two or more assessors should appraise each study. The subjective decisions may be a focus of criticism; thus, we advocate consultation with subject experts and relevant stakeholders when planning your approach. Pragmatic grouping of studies into high, medium and low validity based on simple but discriminatory checklists of "desirable" study features may be necessary if sample sizes are small and do not allow investigation of all the study features individually (for example, Felton et al. 2010, and Isasi-Catalá 2010).

The scope of CEE Systematic Reviews is broad and often interdisciplinary and therefore we seek to be inclusive of different forms of evidence provided their strengths and weaknesses are properly appraised and comparative study weightings are appropriate. However, alongside this inclusivity we expect high levels of transparency providing details of the critical appraisal criteria, how they were applied and the judgements on validity of each study. Normally the full dataset will be provided as an additional supplementary file (see Section 10).

8.3 External validity

External validity is often considered in terms of the relevance of the study; how transferable is it to the context of the question? As noted above, some studies can be of high internal validity (low risk of bias) but may be misleading on account of low external validity (low relevance). A simple example is a high validity study that has been conducted outside the geographical region or in a slightly different ecosystem than the one of interest.

Appraisal of study relevance can be a more subjective exercise than appraisal of study reliability. Estimating the external validity of a study may require the construction of review-specific criteria formed by fit to the question elements or similar subjective measures (see Gough et al. 2012 for examples).

For transparency of reporting, tables of study validity assessment should be included as an appendix or supplementary material. The data validity assessment can be incorporated in narrative synthesis tables if appropriate. Sufficient text should be provided to enable the reader to navigate the tables and understanding the coding and appraisal methods used.

Box 8.1. Examples of Good Practice in Critical Appraisal

The application of critical appraisal of study validity can be broken down into the following four steps in practice;

- establishment of validity assessment criteria;
- deciding on the impact of these criteria on review activities;
- enacting validity assessment and assigning criteria;
- determining the impact of these criteria on review findings.

Example 1. The Importance of Nature to Human Health

In a review of the importance of nature for health, Bowler et al. (2010) undertook critical appraisal of included studies using assessment criteria adapted from a SR of the health literature (specifically, nursing). These criteria included an assessment of: specific methodological bias (e.g. participant self-selection bias), the use of randomisation, the presence of baseline data, and the presence of other confounding variables. Five studies assessed in this review were excluded due to low validity, with the remaining study validity criteria being shown across studies in a bar chart in the results. Finally, study validity weighting was used in sensitivity analyses to compare the results of higher and lower study validity; indicating, for example, that studies with a lower validity score reported a larger effect of nature on tranquillity/calmness than those of higher validity.

Example 2. Peatland Management and Carbon Cycling/GHG Fluxes

In a SR of the impacts of land management activities in lowland peatland ecosystems on greenhouse gas fluxes and carbon cycling, three main experiment types were identified: eddy covariance towers, gas flux chambers, and extractive sampling of soil or soil pore water/air. Alongside an assessment of internal validity (quality) of each study, the external validity (generalisability) of each article was assessed in detail. This process ensured that each relevant article was carefully considered in relation to how well it mapped onto the review question, for example "was it at a comparable spatial or temporal scale?". Subsequently, for internal validity, critical appraisal assessed two types of methodological information. Firstly, general experimental design assessment examined matching of comparator and intervention sites, study season and length, and the time since the intervention occurred. Secondly, specific details relevant to each of the three potential experimental types were assessed; for example, the presence of mitigation measures for trampling around gas flux chambers, the height of eddy covariance towers, and the frequency of sampling. Reasons for possible concern were highlighted during the critical appraisal of each study, and a decision was made as to whether to exclude or include. A validity score based on the presence/absence of bias, appropriateness of controls, precision of methodological design and the presence of confounding variables was given to each study. This score was checked in a subset of studies by a second reviewer and modifications made where necessary. Scores were included as an explanatory variable in meta-analysis as part of a sensitivity analysis to investigate potential differences between studies of higher and lower validity.

Table 8.1. Elements of a data validity assessment of studies included in a SR examining impacts of land management on carbon cycling and greenhouse gas fluxes in boreal and temporal lowland peats.

	Study 1
Methods	Site comparison, GHG flux measured weekly for whole year using closed chambers.
Population	Forested peatlands in Slovenia.
Intervention(s)	Drained plot (19th Century).
Comparator	Undrained plot.
Comparator-matching	Comparator plots close to intervention but distances not disclosed. Soil types moderately different (intervention=rheic hemic histosol (dystric), control=rheic fibric histosol (dystric)).
Outcomes	N2O, CO2, and CH4.
Study design	CI (comparator-intervention).
Level of replication	Plot-level (1 treatment, 1 control), 3 pseudoreplicate samples per plot.
Sampling precision	Weekly measurements 60 minutes each with 3 samples per hour (regression modelling), time=zero measurement.
Confounding variables	Permanent collars account for soil disturbance, foil-covered chambers reduce temperature effects.
Conclusions	Small effective sample size, but good outcome measurement precision. High external validity to SR question. Include in review accounting for low replication.
	Study 2
Methods	Site comparison, GHG flux measured once using closed chambers.
Population	Ombrotrophic fen and minerotrophic bog in Finland.
Intervention(s)	Drained plots (30 years previously).
Comparator	Undrained plots.
Comparator-matching	pH, %N and water table depth measured in all plots and appear similar.
Outcomes	CO2 and CH4.
Study design	CI (comparator-intervention).

Level of replication	Plot-level (one treatment, one control); two regions, one with only ombrotrophic bog, other with ombrotrophic bog and minerotrophic fen. Each site has drained and undrained counterparts. Each site must be treated as a separate study due to substantial differences in plot soil characteristics.
Sampling precision	One sample per plot taken between two and five times over seven month period (exact number unspecified).
Confounding variables	Drained and undrained plots actually only differ very slightly in water table depth, so stated exposure difference may have no real impact. Data extrapolated from very low degree of pseudoreplication (2 to 5 samples over 7 month period).
Conclusions	Drained and undrained plots compared in study but also shown to have minimal differences in water table depth (external validity questionable).

At the end of this stage (if not before) it should become clear what form or forms of synthesis will be possible with the available data. There are a number of different pathways from this point and therefore the following sections become more diverse in terms of the guidance given. They also become more reliant on guiding the reader to more detailed information sources.

Section 9

Data synthesis

Last updated: August 11th 2020

CEE Standards for conduct and reporting

Systematic Reviews (quantitative):

1. **The choice of synthesis method (i.e. narrative synthesis only or with meta-analysis) should be justified in the Protocol on the basis of scoping characteristics of included studies, taking into consideration variability between studies in sample size, study design, context, etc.**
2. **Where meta-analysis is not conducted, a reason for this should be given.**
3. **If meta-analysis is conducted, full details of methods should be presented that justify the approach and enable replication, including study weighting and sensitivity analysis.**
4. **Consideration should be given to study independence and bias (e.g. through sensitivity analysis).**

5. **Effects modifiers (e.g. taxa being considered, location, habitat type, study design etc.) should be investigated statistically through meta-analysis, or descriptively in narrative synthesis.**
6. **Results of critical appraisal should be used in considering individual study findings through statistical or narrative synthesis.**
7. **The narrative synthesis should describe the body of evidence identified using figures and tables that supply information on all eligible studies.**

Systematic Reviews (qualitative):

1. **The choice of synthesis method should be justified in the Protocol on the basis of scoping characteristics of included studies, taking into consideration variability between studies in study design, context, etc.**
2. **The method used to analyse subgroups/subsets of data should be described.**
3. **If all studies were not selected for synthesis an explanation of criteria for selection (e.g. incomplete or missing information) should be provided.**
4. **Consideration should be given to study independence and bias.**
5. **Results of critical appraisal should be used in considering individual study findings through narrative synthesis.**
6. **The narrative synthesis should describe the body of evidence identified using figures and tables that supply information on all eligible studies.**

Systematic Maps:

1. **The choice of mapping and visualisation methods should be justified in the Protocol on the basis of scoping characteristics of included studies, taking into consideration variability between studies.**
2. **The narrative synthesis should describe the body of evidence identified using figures and tables.**
3. **The database of eligible studies with all coded data should be presented in an additional file.**
4. **A map (e.g. geographical or alternative visualisation) should be provided.**

This section includes an overview of different forms of synthesis, narrative, quantitative and qualitative. All Systematic Reviews should present some form of narrative synthesis and many will contain more than one of these approaches (e.g. Bowler et al. 2010). It is not the intention here to give detailed guidelines on synthesis methods since each has its own supporting literature. This Section concentrates on how to make decisions on the correct form of synthesis to conduct.

9.1. Systematic Reviews

9.1.1 Narrative synthesis

Narrative synthesis is the tabulation and/or visualisation (often with descriptive statistics) of the findings of individual primary studies with supporting text to explain the context. A narrative synthesis is often viewed as preparatory when compared with quantitative synthesis and this may be true in terms of application of analytical rigour and lack of statistical power but narrative synthesis has advantages when dealing with broader questions and disparate outcomes. Often narrative synthesis is the only option when faced with a pool of disparate studies of relatively high susceptibility to bias, but such syntheses also accompany quantitative syntheses in order to provide context and background and help characterise the full evidence base. Some form of narrative synthesis should be provided in any Systematic Review, simply to present the context and overview of the evidence. A valuable guide to the conduct of narrative synthesis is provided by Popay (2006).

Narrative synthesis requires the construction of tables, developed from data coding and extraction forms (see Section 8) that provide details of the study or population characteristics, data quality, and relevant outcomes, all of which should have been defined *a priori* in the Protocol. Narrative synthesis should include a statement of the measured effect reported in each study and the Review Team's assessment of study validity (including internal and external validity). Where the validity of studies varies greatly, reviewers may wish to give greater weight to some studies than others. In these instances it is vital that the studies have been subject to standardised a priori critical appraisal with the value judgments regarding both internal and external validity clearly stated. Ideally these will have been subject to stakeholder scrutiny at the Protocol stage. The level of detail employed and emphasis placed on narrative synthesis will be dependent on whether other types of synthesis are also employed. An example of an entirely narrative synthesis (Davies et al. 2006) and a narrative synthesis that complements a quantitative synthesis (Bowler et al. 2010) are available in the CEE Library.

Use of simple vote counting as a form of synthesis (e.g. comparing how many studies showed a positive versus negative or neutral outcome based on statistical significance of the results) should be avoided. Vote counting is misleading because this procedure does not take into account differences in study validity and power. Moreover, vote-counting does not provide an estimate of the magnitude of the effect in question. Whilst tabulation may make it easy for the reader to vote count, the authors should avoid its use in developing and reporting their findings.

Recording of key characteristics of each study included in a narrative synthesis is vital if the Systematic Review is to be useful in summarising the evidence base. Key characteristics are normally presented in tabular form and a minimum list is given below.

- Article reference
- Subject population
- Intervention/exposure variable
- Setting/context
- Outcome measures

- Methodological design
- Relevant reported results

It should be noted here that the interpretation of the results provided by the authors of the study is normally not summarised as this could simply compound subjective assessments or decisions.

9.1.2 Quantitative data synthesis

Usually, when attempting to measure the effect of an intervention or exposure, a quantitative synthesis is desirable. This provides a combined effect and a measure of its variance within and between studies. Quantitative syntheses can be powerful in the sense of enabling the study of the impacts of effect modifiers and increasing power to predict outcomes of interventions or exposures under varying environmental conditions.

Meta-analysis and meta-regression are now commonly used in the environmental sciences and there is a well- developed supporting literature (e.g. Arnqvist & Wooster 1995; Osenberg et al. 1999; Gurevitch & Hedges 2001; Gates 2002; Borenstein et al. 2009; Koricheva et al. 2013) as well as online guidance and training; consequently, we have not provided detailed guidance here. Meta-analysis provides summary effect sizes with each data set weighted according to some measure of its reliability (e.g. with more weight given to large studies with precise effect estimates and less to small studies with imprecise effect estimates). Generally, each study is weighted in proportion to sample size or inverse proportion to the variance of its effect. Meta-regression aims to provide summary effects after adjusting for study-level covariates.

Pooling of individual effects can be undertaken with fixed-effects or random-effects statistical models. Fixed-effects models estimate the combined effect assuming there is a single true underlying effect across the studies, whereas random-effects models assume there is a distribution of effects that depend on study characteristics. Random- effects models include inter-study variability; thus, when there is heterogeneity, a random-effects model usually has wider confidence intervals on its pooled effect than a fixed-effects model (NHS CRD 2001; Khan et al. 2003). Random-or mixed-effects models (containing both random and fixed effects) are often more appropriate for the analysis of ecological data because the numerous complex interactions common in ecology are likely to result in heterogeneity between studies or sites. Exploration of heterogeneity is often more important than the overall pooling from a management perspective, as there is rarely a one-size-fits-all solution to environmental problems.

Relationships between differences in characteristics of individual studies and heterogeneity in results can be investigated as part of the meta-analysis, thus aiding the interpretation of ecological relevance of the findings. Exploration of these differences may be facilitated by construction of tables that group studies with similar characteristics and outcomes together. Important factors that could produce variation in effect size should be defined a priori and their relative importance considered prior to data extraction to make the most efficient use of data. These factors may include differing populations, interventions, outcomes, and methodology. Resulting variation in effect sizes across studies can then be explored by meta-regression.

If sufficient data exist, meta-analyses are often undertaken on subgroups and the significance of differences assessed. Subgroup analyses must be interpreted with caution because statistical power may be limited (Type I errors possible) and multiple analyses of numerous subgroups could result in spurious significance (Type II errors possible). A mixed-effects meta-regression approach might be adopted whereby statistical models including study-level covariates are fitted to the full dataset, with studies weighted according to the precision of the estimate of treatment effect after adjustment for covariates (Sharp 1998).

Despite the attempt to achieve objectivity in reviewing scientific data, considerable subjective judgment is involved when undertaking meta-analyses. These judgements include decisions about choice of effect measure, how data are combined to form datasets, which data sets are relevant and which are methodologically sound enough to be included, methods of meta-analysis, and the issue of whether and how to investigate sources of heterogeneity (Thompson 1994). Reviewers should state these decisions explicitly and distinguish between them to minimise bias and increase transparency.

If possible, a quantitative synthesis should be accompanied by an exploration of possible effects of publication bias. Positive and/or statistically significant results are more readily available than non-significant or negative results because they are more likely published in high-impact journals and in the English language. Whilst searching methodology can reduce this bias, it is still uncertain how influential it might be. There are a number of exploratory plots and tests for publication bias. One example is the funnel plot often accompanied by the Egger (Egger et al. 1997). This approach aims to test for a relationship between the size and precision of study effects, plotted on x- and y-axis of the funnel plot. However, a funnel plot can change greatly depending on the scale of the precision (Lau et al. 2006), and only the trial size is appropriate for effect measures used in ecology, such as the standardised mean difference (SMD) or response ratio. Another approach is to calculate the fail safe number, which is the number of null result studies that would have to be added to a meta-analysis to lower the significance or the magnitude of the effect to a specified level (e.g. where it would be considered statistically or biologically non-significant), but see Scargle (2000). Wherever possible, grey literature and unpublished studies should be included in a meta-analysis to allow direct assessment of publication bias by comparison of effect sizes in published and unpublished studies.

9.1.3 Qualitative data synthesis

It is common in the social sciences to employ qualitative methods where the views of individual people are recorded in relation to a question. When open ended question are asked and complex answers received, the data are not formally quantified. In such studies the authors are often seeking to characterise the range of views or reactions to a particular question or set of questions. The role of qualitative data synthesis is therefore quite distinct and serves to increase understanding of some environmental issues. Evidence from qualitative studies may help and generate hypotheses that might be further tested by quantitative methods. Qualitative data evidence may also complement quantitative and contribute to a mixed method approach together with quantitative data.

A synthesis of evidence from qualitative research can explore questions such as how do people experience interventions that impact on their environment or social settings (for example, moving from a farming system that uses non-organic pest control to one that employs integrated or organic pest control); why does an intervention work (or not), for whom and in what circumstances? It may be desirable to draw on qualitative evidence to address questions such as what are the barriers and facilitators to accessing given interventions, or what impact do specific barriers and facilitators have on people, their experiences and behaviours

To date qualitative methods have been applied only rarely in environmental evidence synthesis. An example of the use of qualitative synthesis in a CEE review can be found in Pullin et al. (2013). Further information on these methods can be found in Gough et al. (2012) and Noyes et al. (2011).

9.2 Systematic Maps

9.2.1 Narrative synthesis

Narrative synthesis in Systematic Maps is the tabulation and/or visualisation (often with descriptive statistics) of the characteristics of individual primary studies with supporting text to explain the context. Some form of narrative synthesis should be provided in any Systematic Map, simply to present the context and overview of the distribution and abundance of evidence.

Narrative synthesis requires the construction of tables, developed from data coding forms (see Section 8) that provide details of the study or population characteristics and relevant outcomes types, all of which should have been defined *a priori* in the Protocol. Narrative synthesis should not include a statement of the measured effect reported in each study but the Review Team's assessment of study validity (including internal and external validity) may be included. The level of detail employed and emphasis placed on narrative synthesis will be dependent on the form of mapping and data visualisation employed.

9.2.2 Mapping and data visualisation

The process of mapping and presentation of data can take many forms and this guidance does not wish to be overly prescriptive in what is a fast moving field (see James et al 2016 for a detailed discussion of methodologies for the production of Systematic Maps). Presentation of maps can range from a simple spreadsheet format to innovative forms of data visualisation that make the evidence base easier to interrogate and extract information of interest to the user. Good examples of data visualisation are McKinnon et al. (2016) and Haddaway et al. (2014).

Recording of key characteristics of each study included in a narrative synthesis is vital if the Systematic Map is to be useful in summarising the evidence base. Key characteristics stated in the Protocol must be fully presented in at least tabular form. Below is a minimum list of characteristics that will normally be enhanced through data coding of other variables of interest.

- Article reference
- Subject population
- Intervention/exposure variable

- Setting/context
- Outcome measures
- Methodological design

Section 10

Interpreting findings and reporting conduct

Last updated September 3rd 2020

10.1 The interpretation of evidence syntheses

CEE Evidence synthesis methodologies seek to collate and synthesise data in order to present reliable evidence in relation to the review question. The strength of the evidence base and implications of the results for decision-making require careful consideration and interpretation. The discussion and conclusions may consider the implications of the evidence in relation to practical decisions, but the decision-making context may vary, leading to different decisions based on the same evidence. Authors should, where appropriate, explicitly acknowledge the variation in possible interpretation and simply present the evidence so as to inform rather than offer advice. Recommendations that depend on assumptions about resources and values should be avoided (Khan et al. 2003, Deeks et al. 2005).

Deeks et al (2005) offer the following advice that is of relevance here. Authors and end-users should be wary of the pitfalls surrounding inconclusive evidence and should beware of unwittingly introducing bias in their desire to draw conclusions rather than pointing out the limits of current knowledge. Where reviews are inconclusive because there is insufficient evidence, **it is important not to confuse 'no evidence of an effect' with 'evidence of no effect'**. The former may not provide a basis for change to existing policy or practice, but has an important bearing on future research, whereas the latter could have considerable ramifications for current policy or practice.

Review authors, and to a lesser extent end-users, may be tempted to reach conclusions that go beyond the evidence that is reviewed or to present only some of the results. Authors must be careful to be balanced when reporting on and interpreting results. For example, if a 'positive' but statistically non-significant trend is described as 'promising', then a 'negative' effect of the same magnitude should be described as a 'warning sign'. Other examples of unbalanced reporting include one-sided reporting of sensitivity analyses or explaining non-significant positive results but not negative ones. If the confidence interval for the estimate of difference in the effects of interventions overlaps the null value, the analysis is compatible with both a true beneficial effect and a true harmful effect. If one of the possibilities is mentioned in the conclusion, the other possibility should be mentioned as well and both should be given equal consideration in discussion of results. One-sided attempts to explain results with reference to indirect evidence external to the review should be avoided. Considering results in a blinded manner can avoid these pitfalls (Deeks *et al.* 2005). Authors should consider how the results would be presented and framed in the conclusions and discussion if the direction of the results was reversed.

10.1.1 Limitations of an evidence synthesis

Biases can occur in the evidence synthesis process, which do not impair the raw data themselves but may affect the findings of the synthesis (through a biased sample of articles) and should be fully considered and reported (see review in Borenstein et al. 2009). For example:

Publication bias: statistically significant results are more prone to be published than non significant ones. Yet, there is no strict relationship between the quality of the methodology and the significance of results, and thus, their publication. A good methodology may lead to non significant results and be kept as a grey article.

Language bias: searching is generally undertaken in English because it is the most common language used in scientific writing. This may result in an over-representation of statistically significant results (Egger et al. 1997; Jüni et al. 2002) because they are more likely to be accepted in the English scientific literature.

Availability bias: only the studies that are easily available are included in the analysis, whilst other significant results may exist but are less easily available (this can be an increasing problem as many private companies have their own research teams and publish their own articles or reports). Similarly, a **confidentiality bias** may exist in some sensitive topics (eg GMO, nuclear power) because some research results may not be available for security reasons.

Cost bias: time and resources necessary for a thorough search are not always available, which could lead to the selection of the studies only available free or at low cost.

Familiarity bias: the researcher limited the search to articles relevant to his/her own discipline.

Duplication bias: some studies with statistically significant results may be published more than once (Tramer et al. 1997).

Citation bias: Studies with significant results are more likely to be cited by other authors and thus easier to be found during the search (Gøtzsche 1997; Ravnskov 1992).

All these biases can be considered when reporting 'limitations of the evidence synthesis' and several methods exist to quantify their impacts on the results (Borenstein et al. 2009).

10.2 Reporting conduct of evidence syntheses

CEE standards require a high level of reporting of the conduct of evidence syntheses so as to ensure high transparency and repeatability allowing others to test replicability of findings.

Each of the conduct Section (5-9) has guidance on reporting. In addition CEE now recommends using the RepOrting standards for Systematic Evidence Syntheses (ROSES) checklist (Haddaway et al. in press) as this will be used by editors and peer reviewers when appraising reports.

10.3 Reporting findings of evidence syntheses

Evidence Syntheses are most often conducted to assess available evidence of effectiveness or of impact. In so doing, Systematic Reviews (not SMs) assess the strength of a causal inference (Hill 1971). Aspects that may be reported in the conclusion section include:

1. The quality/reliability of the included studies.
2. The relevance/external validity of the included studies.
3. The size and statistical significance of the observed effects.
4. The consistency of the effects across studies or sites and the extent to which this can be explained by other variables (effect modifiers).
5. The clarity of the relationship between the intensity of the intervention and the outcome.
6. The existence of any indirect evidence that supports or refutes the inference.
7. The lack of other plausible competing explanations of the observed effects (bias or confounding).

In a review concerning the impacts of liming streams and rivers on fish and invertebrates, Mant et al. (2011) discuss all of the above points in a good example of Systematic Review conclusions. In addition to discussing the limitations of their review, the authors describe the range of quality of studies included, the size and consistency of the effect observed across studies, the link between intervention intensity and outcome, the presence of effect modifiers, the presence of evidence in support/refute of the review findings, and the potential for other causative factors for the observed effects.

There is a range of approaches to grading the strength of evidence presented in health-related reviews, but there is no universal approach (Deeks et al. 2005). We suggest that authors of reviews in environmental management explicitly state weaknesses associated with each of the aspects above, but the overall impact they make on conclusions can only be considered subjectively.

10.3.1 Implications for policy and practice

A key objective of Systematic Review is to inform decision-makers of the implications of the best available evidence relating to a question of concern, and enable them to place this evidence in context, in order to make a decision on the best course of action. Providing evidence that increases capacity to predict the outcomes of alternative actions should lead to better decision making.

End-users will need to decide, either implicitly or explicitly, how applicable the evidence presented in a Systematic Review is to their particular circumstances (Deeks et al. 2005). This is particularly critical in environmental management where many factors may vary between sites and it seems likely that many interventions/actions will vary in their effectiveness/impact depending on a wide range of potential environmental variables. Authors should therefore highlight where the evidence is likely to be applicable and equally importantly where it may not

be applicable with reference to variation between studies and study characteristics (see 8.3 External validity).

Clearly, variation in the ecological context and geographical location of studies can limit the applicability of results. Authors should be aware of the timescale of included studies, which may be insufficiently short to make long-term predictions. Variation in application of the intervention may also be important (and difficult to predict), but authors should be aware of differences between *ex situ* and *in situ* treatments (measuring efficacy versus effectiveness respectively) where they are combined and should also consider the implications of applying the same intervention at different scales. Variation in baseline risk may also be an important consideration in determining the applicability of results, as the net benefit of any intervention depends on the risk of adverse outcomes without intervention, as well as on the effectiveness of the intervention (Deeks et al. 2005).

Where review authors identify predictable variation in the relative effect of the intervention or exposure in relation to the specified reasons for heterogeneity, these should be highlighted. However, these relationships require cautious interpretation (because they are only correlations), particularly where sample sizes are small, data points are not fully independent and multiple confounding occurs. **When reporting implications of the review findings for policy and practice, the emphasis should be on objective information and not on subjective advocacy.**

10.3.2 Implications for research

Rather like primary scientific studies, most Systematic Reviews will generate more questions than they answer. Knowledge gaps will be frequent, as will areas where the quality of science conducted to date is inadequate. In conducting an Systematic Review, critically appraising the quality of existing studies and attempting to assess the available evidence in terms of its fitness for purpose, reviewers should be able to draw conclusions concerning the need for further research. This need may simply be reported in the form of knowledge gaps but may often consist of recommendations for the design of future studies that will generate data of sufficient quality to improve the evidence base and decrease the uncertainty surrounding the question.

10.4 Format for CEE Reports

The format for submitting full reports can be found on the Environmental Evidence website by following the links below

For **Systematic Maps**

For **Systematic Reviews**

10.4.1 Additional files

To maximise transparency Systematic Reviews should normally be supported by a number of supplementary materials made available in additional files linked from the main text. Authors should ROSES for guidance on what should be reported. The following is a minimal list of

expected information (note other additional files may be provided depending on the size and complexity of the synthesis);

References

1. A report of literature scoping containing combinations of search strings and the outcome of searches of different databases (this is usually as an appendix with the Protocol).
2. A list of articles excluded after reading the full text, including reasons for exclusion (note: a list of articles included is expected in the main text).
3. A list of articles that could not be obtained at full text: such articles are therefore potentially relevant but not fully screened.
4. Data extraction and validity assessment tables for Systematic Reviews or data coding tables for Systematic Maps; for example Excel files with data extracted from each included study (this may be included in the main text if a small number of studies is included or may be provided in several files for larger Systematic Reviews).
5. *Last updated January 7th, 2018*

6. AIASSA, E., HIGGINS, J. P. T., FRAMPTON, G. K., GREINER, M., AFONSO, A., AMZAL, B., DEEKS, J., DORNE, J. L., GLANVILLE, J., LÖVEI, G. L., NIENSTEDT, K., O'CONNOR, A. M., PULLIN, A. S., RAJIĆ, A. & VERLOO, D. 2015. Applicability and feasibility of systematic review for performing evidence-based risk assessment in food and feed safety. *Critical Reviews in Food Science and Nutrition* 55**,** 1026-1034.
7. ALTMAN, D.G. 1991. Measuring agreement. In: Altman D.G. (Ed.), Practical statistics for medical research. London: Chapman and Hall.
8. ARNQVIST, G. & WOOSTER, D. 1995. Meta-analysis: synthesizing research findings in ecology and evolution. *Trends in Ecology & Evolution,* 10**,** 236-240.
9. AVENELL, A., HANDOLL, H. & GRANT, A. 2001. Lessons for search strategies from a systematic review, in The Cochrane Library, of nutritional supplementation trials in patients after hip fracture. *American Journal of Clinical Nutrition,* 73**,** 505-510.
10. BAILEY, B. J. 2002. Duplicate publication in the field of otolaryngology-head and neck surgery. *Otolaryngology Head and Neck Surgery,* 126**,** 211-216.
11. BARDEN, J., EDWARDS, J. E., MCQUAY, H. J. & MOORE, R. A. 2003. Oral valdecoxib and injected parecoxib for acute postoperative pain: a quantitative systematic review. *BMC Anesthesiology,* 3**,** 1-1.
12. BAYLISS, H. R. & BEYER, F. R. 2015. Information retrieval for ecological syntheses. *Research Synthesis Methods,* 6**,**136-148.
13. BAYLISS, H. R., HADDAWAY, N. R., EALES, J., FRAMPTON, G. K. & JAMES, K. L. 2016. Updating and amending systematic reviews and systematic maps in environmental management. *Environmental Evidence,* 5**,** 20.
14. BAYLISS, H. R., SCHINDLER, S., ADAM, M., ESSL, F. & RABITSCH, W. 2017. *Environmental Evidence*, **6**:21 https://doi.org/10.1186/s13750-017-0100-4.
15. BERNARD, H.R., 2006. Research Methods in Anthropology: Qualitative and Quantitative Approaches. Altamira Press, Lanham, New York, Toronto, Oxford: Rowman & Littlefield Publishers, Inc.

16. BERNES, C., CARPENTER, S. R., GÅRDMARK, A., LARSSON, P., PERSSON, L., SKOV, C., SPEED, J. D. M. & VAN DONK, E. 2015. What is the influence of a reduction of planktivorous and benthivorous fish on water quality in temperate eutrophic lakes? A systematic review. *Environmental Evidence* 4:7
17. BERO, L., ANGLEMYER, A., VESTERINEN, H. & KRAUTH, D. 2016. The relationship between study sponsorship, risks of bias, and research outcomes in atrazine exposure studies conducted in non-human animals: systematic review and meta-analysis. *Environment International.* 92-93, 597-604.
18. BOOTH, A. 2004. Formulating answerable questions. *In:* Booth, A. & Brice, A. (Eds.) *Evidence-based practice: an information professional's handbook.*London: Facet.
19. BOOTH, A. 2010. How much searching is enough? Comprehensive versus optimal retrieval for technology assessments. *International Journal of Technology Assessment in Health Care*. 26(4), 431-435.
20. BORENSTEIN, M., HEDGES, L. V., HIGGINS, J. P. T. & ROTHSTEIN, H. R. 2009. *Introduction to Meta-Analysis*, John Wiley and Sons.
21. BOTTRILL, M., CHENG, S., GARSIDE, R., WONGBUSARAKUM, S., ROE, D., HOLLAND, M. B., EDMOND, J. & TURNER, W. R. 2014. What are the impacts of nature conservation interventions on human well-being: a systematic map protocol. *Environmental Evidence* 3**,** 16.
22. BOWLER, D., BUYUNG-ALI, L., KNIGHT, T., & PULLIN, A .S. 2009. The importance of nature for health: is there a specific benefit of contact with green space? *Environmental Evidence*: www.environmentalevidence.org/SR40.htm
23. BOWLER, D., BUYUNG-ALI L., KNIGHT, T. & PULLIN A. S. 2010. How effective is 'greening' of urban areas in reducing human exposure to ground level ozone concentrations, UV exposure and the 'urban heat island effect'? Environmental Evidence: www.environmentalevidence.org/SR41.html
24. BUSSELL, J., JONES, D. L., HEALEY, J. R. & PULLIN, A. S. 2010. How do draining and re-wetting affect carbon stores and greenhouse gas fluxes in peatland soils? CEE review 08-012 (SR49). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR49.html.
25. BRAMER, W. M., GIUSTINI, D., KRAMER, B. M. R. & ANDERSON, P. F. 2013. The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic review: a review of searches used in systematic reviews. *Systematic Reviews* 2**,** 115.
26. CAMPBELL COLLABORATION, STEERING GROUP OF THE CAMPBELL COLLABORATION 2014. Campbell Systematic Reviews: Policies and Guidelines. Campbell Policies and Guidelines Series No. 1. DOI:10.4073/cpg.2016.1
27. CENTRE FOR REVIEWS AND DISSEMINATION (CRD). 2009. Systematic Reviews. CRD's Guidance for Undertaking Reviews in Health Care. York: CRD, University of York.
28. CHAN, A., HRÓBJARTSSON, A., HAAHR, M., GØTZSCHE, .P. & ALTMAN, D. 2004. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*;291,2457-65.
29. CHOI, W. S., SONG, S. W., OCK, S. M., KIM, C. M., LEE, J., CHANG, W. J. & KIM, S. H. 2014. Duplicate publication of articles used in meta-analysis in Korea. *SpringerPlus,* 3**,** 182.

30. COLLABORATION FOR ENVIRONMENTAL EVIDENCE (CEE). 2013. Guidelines for systematic review and evidence synthesis in environmental management. Version 4.2. https://environmentalevidence.org/wp-content/uploads/2014/06/Review-guidelines-version-4.2-final.pdf
31. CORLETT, R. T. 2011. Trouble with the gray literature. *Biotropica,* 43**,** 3-5.
32. CÔTÉ, I. M., MOSQUEIRA, I. & REYNOLDS, J. D. 2001. Effects of marine reserve characteristics on the protection of fish populations: A meta-analysis. *Journal of Fish Biology,* 59**,** 178-189.
33. DAVIES, Z. G., TYLER, C., STEWART, G. B. & PULLIN, A. S. 2006. Are current management recommendations for conserving saproxylic invertebrates effective? CEE review 05-011 (SR17). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR17.html.
34. DEEKS, J. J., HIGGINS, J. P. T. & ALTMAN, D. G. 2005. Analysing and presenting results. In: Higgins, J.P.T & Green, S. (Eds) Cochrane Handbook for Systematic Reviews of Interventions 4.2.5 [updated May 2005]; Section 8. (www.cochrane.org/resources/handbook/hbook.htm).
35. DICKERSIN, K. 2005. Publication bias: recognizing the problem, understanding its origins and scope, and preventing harm. *In:* Rothstein, H. R., Sutton, A. J. & Borenstein, M. (Eds.) *Publication bias in meta-analysis: prevention, assessment, and adjustments.* London: Wiley.
36. DEPARTMENT FOR ENVIRONMENT FOOD AND RURAL AFFAIRS (DEFRA). 2015. Emerging tools and techniques to deliver timely and cost effective evidence reviews. Final Report WT1552. London, DEFRA.
37. DOERR, E. D., DORROUGH, J., DAVIES, M. J., DOERR, V. A. J. & MCINTYRE, S. 2015. Maximising the value of systematic reviews in ecology when data or resources are limited. *Austral Ecology,* 40**,** 1-11.
38. DOWNING, J. A., OSENBERG, C. W. & SARNELLE, O. 1999. Meta-analysis of marine nutrient-enrichment experiments: Variation in the magnitude of nutrient limitation. *Ecology,* 80**,** 1157-1167.
39. DUFFY, L., DUNLAP, K., GODDUHN, A. 2014. Bias, complexity, and uncertainty in ecosystem risk assessment: pharmaceuticals, a new challenge in scale and perspective. *Environmental Research Letters*. 9,091004,1-3.
40. DUFFIELD, S., AEBISCHER, N. 1994. The effect of spatial scale of treatment with dimethoate on invertebrate population recovery in winter wheat. *Journal of Applied Ecology*. 31,263-281.
41. EFSA (European Food and Safety Authority). 2010. Application of systematic review methodology to food and safety assessments to support decision making. EFSA Journal. 8(6), 1637.
42. EDWARDS, P., CLARKE, M., DIGUISEPPI, C., PRATAP, S., ROBERTS, I. & WENTZ, R. 2002. Identification of randomized controlled trials in systematic reviews: Accuracy and reliability of screening records. *Statistics in Medicine,*21**,** 1635-1640.
43. EGGER, M., DAVEY-SMITH, G., SCHNEIDER, M., & MINDER, C. 1997. Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315, 629-634.
44. ENGLUND, G., SARNELLE, O. & COOPER, S. D. 1999. The importance of data-selection criteria: meta-analyses of stream predation experiments. *Ecology,* 80**,**1132-1141.

45. FAZEY, I., SAILSBURY, J. G., LINDENMAYER, D. B., MAINDONALD, J. & DOUGLAS, R. 2004. Can methods applied in medicine be used to summarize and disseminate conservation research? *Environmental Conservation,* 31,190-198.
46. FELTON, A., KNIGHT, E., WOOD, J., ZAMMIT, C.& LINDENMAYER, D. B. 2010. A meta-analysis of fauna and flora species richness and abundance in plantations and pasture lands. CEE review 09-012 (SR73). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR73.html.
47. FRAMPTON, G. K., LIVOREIL, B. & PETROKOFSKY, G. 2017. Eligibility screening in evidence synthesis of environmental management topics. *Environmental Evidence*, 6, 27.
48. GATES, S. 2002. Review of methodology of quantitative reviews using meta-analysis in ecology. *Journal of Animal Ecology,* 71**,** 547-557.
49. GERBER, A. S. & GREEN, D. P. 2012. *Field Experiments: Design, Analysis, and Interpretation*. W.W. Norton & Co.
50. GIBSON, G. & RUSSELL, I. 2006. Flying in Tune: Sexual Recognition in Mosquitoes. *Current Biology* **16** (13),1311-1316.
51. GIUSTINI, D. & BOULOS, M. N. K. 2013. Google Scholar is not enough to be used alone for systematic reviews. *Online Journal of Public Health Informatics,* 5**,**1-9.
52. GLANVILLE, J. Searching bibliographic databases. [In press] In: Cooper, H. C., Hedges, L. V. & J. C. Valentine, C. (Eds). The Handbook of Research Synthesis and Meta-Analysis, 3rd edition. New York, NY: Russell Sage Foundation.
53. GLASS, G. 1976. Primary, secondary, and meta-analysis of research. *Educational Researcher*; 5, 3-8.
54. GØTZSCHE, P. C. 1989. Multiple publication of reports of drug trials. *European Journal of Clinical Pharmacology* 36**,**429-432.
55. GØTZSCHE, P. C. & IOANNIDIS, J. P. A. 2012. Content area experts as authors: helpful or harmful for systematic reviews and meta-analyses? *BMJ : British Medical Journal,* 345**,** 4 pp.
56. GOUGH, D., OLIVER, S. & THOMAS, J. 2012. *An introduction to systematic reviews,* London, Sage Publications Ltd.
57. GRINDLAY, D. J. C., BRENNAN, M. L. & DEAN, R. S. 2012. Searching the veterinary literature: a comparison of the coverage of veterinary journals by nine bibliographic databases. *Journal of Veterinary Medicine Education,* 39**,** 404-412.
58. GUREVITCH, J. & HEDGES, L. V. 1999. Statistical issues in ecological meta-analyses. *Ecology,* 80**,** 1142–1149.
59. GUREVITCH J. & HEDGES L. V. 2001. Meta-analysis Combining the results of independent experiments. In: Scheiner, S. M. & J. Gurevitch, J. (Eds) Design and Analysis of Ecological Experiments. pp. 347-369. Oxford University Press, New York.
60. HADDAWAY, N. R. 2015. The use of web-scraping software in searching for grey literature. *The Grey Journal,* 11**,** 186-190.
61. HADDAWAY, N. R. & BAYLISS, H. R. 2015. Shades of grey: two forms of grey literature important for reviews in conservation. *Biological Conservation,* 191**,**827-829.
62. HADDAWAY, N. R., COLLINS, A. M., COUGHLIN, D. & KIRK, S. 2015. The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PLoS ONE,* 10**,** e0138237.

63. HADDAWAY, N. R., COLLINS, A. M., COUGHLIN, D. & KIRK, S. 2017. A rapid method to increase transparency and efficiency in web-based searches. *Environmental Evidence,* 6**,** 1.
64. HADDAWAY, N.R., STYLES, D & PULLIN, A.S. 2014. Evidence on the environmental impacts of farm land abandonment in high altitude/mountain regions: a systematic map. *Environmental Evidence* 3, 17.
65. HIGGINS, J. P. T., ALTMAN, D., GØTZSCHE. P., JÜNI, P., MOHER, D., OXMAN, A., et al. 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. 343,d5928,1-9.
66. HIGGINS, J.P.T., DEEKS, J.J. 2011. Selecting studies and collecting data. In: Higgins, J. P. T. & Green, S, (Eds). Cochrane Handbook for Systematic Reviews of Interventions. Chichester: Wiley-Blackwell; pp. 151-86.
67. HIGGINS, J. P. T. & GREEN, S. 2011. *Cochrane handbook for systematic reviews of interventions,* Chichester, Wiley.
68. HILL, A. B. 1971. Principles of medical statistics. *Lancet* 9: 312-20.
69. HOLMAN, L., HEAD, M., LANFEAR, R. & JENNIONS, M. 2015. Evidence of experimental bias in the life sciences: why we need blind data recording. *PLoS Biology*. 13(7),e1002190,1-12.
70. HOLST, R. & FUNK, C. J. 2005. State of the art of expert searching: results of a Medical Library association survey. *Journal of the Medical Library Association,* 93**,** 45-52.
71. HOPEWELL, S., MCDONALD, S., CLARKE, M. J. & EGGER, M. 2007. Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database of Systematic Reviews,* Issue 2.
72. IOANNIDIS, J. P. A. 2005. Why Most Published Research Findings Are False. *PLoS Med* 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124
73. JAMES, K. L., RANDALL, N. P. & HADDAWAY, N. R. 2016. A methodology for systematic mapping in environmental sciences. *Environmental Evidence,* 5**,** 7.
74. JOHNSON, V., FITZPATRICK, I., FLOYD, R. & SIMMS, A. 2011. What is the evidence that scarcity and shocks in freshwater resources cause conflict instead of promoting collaboration? CEE review 10-010. Collaboration for Environmental Evidence: www.environmentalevidence.org/SR10010.html.
75. KHAN, K. S., KUNZ, R., KLEIJNEN, J. & ANTES, G. 2003. *Systematic reviews to support evidence-based medicine: how to apply findings of healthcare research*. Royal Society of Medicine Press Ltd, London.
76. KHORSAN, R., CRAWFORD, C. 2014. External validity and model validity: a conceptual approach for systematic review methodology. *Evidence-Based Complementary and Alternative Medicine.* Volume 2014, Article ID 694804,1-12.
77. KIRKHAM, J., DWAN, K., ALTMAN, D., GAMBLE, C., DODD, S., SMYTH, R., et al. 2010. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*., 340, c365,1-10.
78. KORICHEVA, J, GUREVITCH, J & MENGERSEN, K. 2013.*Handbook of meta-analysis in ecology and evolution*. Princeton University Press.
79. KUGLEY, S., WADE, A., THOMAS, J., MAHOOD, Q., KLINT-JØRGENSEN, A. M., HAMMERSTRØM, K. & SATHE, N. 2016. Searching for studies: a guide to information retrieval for Campbell Systematic Reviews. Version 1.1. Campbell Methods Series. Method Guide 1. Oslo: The Campbell Collaboration.

80. LAND, M., GRANÉLI, W., GRIMWALL, A., HOFFMANN, C. C., MITSCH, W. J., TONDERSKI, K. S. & VERHOEVEN, J. T. A. 2013. How effective are created or restored freshwater wetlands for nitrogen and phosphorus removal? A systematic review protocol. *Environmental Evidence,* 2**,** 16.
81. LANGER, L., ERASMUS, Y., TANNOUS, N. & STEWART, R. 2017. How stakeholder engagement has led us to reconsider definitions of rigour in systematic reviews. *Environmental Evidence,* 6**,** 20.
82. LAU, J., IOANNIDIS, J. P. A., TERRIN, N, SCHMID, CH, & OLKIN, I. 2006. The case of the misleading funnel plot. *BMJ* 333: 597–600.
83. LEIMU, R. & KORICHEVA, J. 2004. Cumulative meta-analysis: a new tool for detection of temporal trends and publication bias in ecology. *Proceedings of the Royal Society B: Biological Sciences,* 271**,** 1961-1966.
84. LEIMU, R. & KORICHEVA, J. 2005. What determines the citation frequency of ecological papers? *Trends in Ecology and Evolution,* 20**,** 28-32.
85. LIBERATI, A., ALTMAN, D. G., TETZLAFF, J., MULROW, C., GØTZSCHE, P. C., IOANNIDIS, J. P. A., CLARKE, M., DEVEREAUX, P. J., KLEIJNEN, J. & MOHER, D. 2009. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Medicine,* 6**,** e1000100.
86. LIPSEY M. W. & WILSON D. B. 2001. Practical Meta-analysis. Applied Social Research Methods Series. Volume 49. Sage Publications, Thousand Oaks, California.
87. LIVOREIL, B., GLANVILLE, J., HADDAWAY, N. R., BAYLISS, H., BETHEL, A., DE LACHAPELLE, F. F., ROBALINO, S., SAVILAAKSO, S., ZHOU, W., PETROKOFSKY, G. & FRAMPTON, G. 2017. Systematic searching for environmental evidence using multiple tools and sources. *Environmental Evidence,* 6**,** 23.
88. LORTIE, C. J., AARSSEN, L. W., BUDDEN, A. E., KORICHEVA, J. K., LEIMU, R. & TREGENZA, T. 2007. Publication bias and merit in ecology. *Oikos,* 116**,** 1247-1253.
89. MACKENZIE, R. S., MCMANUS, C., HARRISON, V., MASON, O. 2016. Reflections on the process of using systematic review techniques to evaluate the literature regarding the neurotoxicity of low level exposure to organophosphate pesticides. *Environment International*. 92-93:569-573.
90. MCDONAGH, M., PETERSON, K., RAINA, P., CHANG, S. & SHEKELLE, P. 2013. Avoiding Bias in Selecting Studies. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville (MD), AHRQ Publication No. 13-EH045-EF. Agency for Healthcare Research and Quality (US)
91. MANT, R., JONES, D., REYNOLDS, B., ORMEROD, S. & PULLIN A. S. 2011. What is the impact of liming of streams and rivers on the abundance and diversity of fish and invertebrates? CEE review 09-015 (SR76). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR76.html.
92. MAHOOD, Q., EERD, D. & IRVIN, E. 2014. Searching for grey literature for systematic reviews: challenges and benefits. *Research Synthesis Methods,* 3**,** 221-234.
93. MCKINNON, M. C., CHENG, S. H., DUPRE, S., EDMOND, J., GARSIDE, R., GLEW, L., HOLLAND, M. B., LEVINE, E., MASUDA, Y. J., MILLER, D. C., OLIVEIRA, I., REVENAZ, J., ROE, D., SHAMER, S., WILKIE, D., WONGBUSARAKUM, S. & WOODHOUSE, E. 2016. What are the effects of nature conservation on human well-

being? A systematic map of empirical evidence from developing countries. *Environmental Evidence,* 5**,** 8.
94. MOHER, D., JADAD, A. R., NICHOL, G., PENMAN, M., TUGWELL, P. & WALSH, S. 1995. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clinical Trials,* 16**,** 62-73.
95. MULLINS, M. M., DELUCA, J. B., CREPAZ, N. & LYLES, C. M. 2014. Reporting quality of search methods in systematic reviews of HIV behavioural interventions (2000–2010); are the searches clearly explained, systematic and reproducible? *Research Synthesis Methods,* 5**,** 116-130.
96. NHS CENTRE FOR REVIEWS AND DISSEMINATION. 2001. *Undertaking systematic review of research on effectiveness*. NHS CRD, University of York.
97. NOYES, J., POPAY, J., PEARSON, A., HANNES, K. & BOOTH, A. 2011. Chapter 20: Qualitative research and Cochrane reviews. In: Higgins, J. P. T. & Green, S. (Eds), Cochrane Handbook for Systematic Reviews of Interventions Version 5.1. The Cochrane Collaboration. www.handbook.cochrane.org.
98. OJANEN, M., MILLER, D., ZHOU, W., MSHALE, B., MWANGI, E. & PETROKOFSKY, G. 2014. What are the environmental impacts of property rights regimes in forests, fisheries and rangelands? A systematic review protocol. *Environmental Evidence,* 3**,** 19.
99. OJANEN, M., ZHOU, W., MILLER, D. C., NIETO, S. H., MSHALE, B. & PETROKOFSKY, G. 2017. What are the environmental impacts of property rights regimes in forests, fisheries and rangelands? *Environmental Evidence,* 6**,** 12.
100. O'LEARY, B. C., KVIST, K., BAYLISS, H. R., DERROIRE, G., HEALEY, J. R., HUGHES, K, KLEINSCHROTH, F., SCIBERRAS, M., WOODCOCK, P. & PULLIN, A. S. 2016. The reliability of evidence review methodology in environmental science and conservation. *Environmental Science and Policy* 64, 75-82.
101. ORTEGA, J. L. 2014. Academic search engines: a quantitative outlook. *Online Information Review,* 39**,** 435-436.
102. OSENBERG, C. W., SARNELLE, O., COOPER, S. D. & HOLT, R. D. 1999. Resolving ecological questions through meta-analysis: Goals, metrics, and models. *Ecology,* 80**,** 1105-1117.
103. PAGE, M., HIGGINS, J. 2016. Rethinking the assessment of risk of bias due to selective reporting: a cross-sectional study. *Systematic Reviews*. 5,108,1-8.
104. PEAT, J. 2001. *Health sciences research: A handbook of quantitative methods*. London: SAGE Publishers Ltd., 328 pp.
105. PETERSEN, K. & ALI, N. B. 2011. Identifying Strategies for Study Selection in Systematic Reviews and Maps. *International Symposium on Empirical Software Engineering and Measurement.* Banff, AB, Canada
106. PETTICREW, M. & ROBERTS, H. 2006. *Systematic reviews in the social sciences. A practical guide,* Oxford, Blackwell.
107. POPAY, J. 2006. *Moving Beyond Effectiveness. Methodological issues in the synthesis of diverse sources of evidence*. National Institute for Health and Clinical Evidence, UK.
108. PRIESNITZ, K. U., VAASEN, A. & GATHMANN, A. 2016. Baseline susceptibility of different European lepidopteran and coleopteran pests to Bt proteins expressed in Bt Maize: a systematic review. *Environmental Evidence,* 5**,** 27.

109. PULLIN, A. S., BANGPAN, M., DALRYMPLE, S. E., DICKSON, K., HADDAWAY, N. R., HEALEY, J. R., HAUARI, H., HOCKLEY, N., JONES, J. P. G., KNIGHT, T. M., VIGURS, C. & OLIVER, S. 2013. Human Well-Being Impacts of Terrestrial Protected Areas. *Environmental Evidence* 2:19.
110. PULLIN, A. S. & KNIGHT, T. M. 2003. Support for decision making in conservation practice: an evidence-based approach. *Journal for Nature Conservation,*11**,** 83-90.
111. PULLIN, A. S., KNIGHT, T. M. & WATKINSON, A. R. 2009. Linking reductionist science and holistic policy using systematic reviews: unpacking environmental policy questions to construct an evidence-based framework. *Journal of Applied Ecology,* 46**,** 970-975.
112. RADER, T., MANN, M., STANSFIELD, C., COOPER, C. & SAMPSON, M. 2014. Methods for documenting systematic review searches: a discussion of common issues. *Research Synthesis Methods,* 5**,** 98-115.
113. RANDALL, N. P. & JAMES, K. L. 2012. The effectiveness of integrated farm management, organic farming and agri-environment schemes for conserving biodiversity in temperate Europe – A systematic map. *Environmental Evidence,* 1.
114. RATHBONE, J., HOFFMANN, T. & GLASZIOU, P. 2015. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic Reviews,* 4**,** 80.
115. ROBERTS, P. D., STEWART, G. B. & PULLIN, A. S. 2006. Are review articles a reliable source of evidence to support conservation and environmental management? A comparison with medicine. *Biological Conservation,* 132**,** 409-423.
116. RODRÍGUEZ, L., HOGARTH, N. J., ZHOU, W., XIE, C., ZHANG, K. & PUTZEL, L. 2016. China's conversion of cropland to forest program: a systematic review of the environmental and socioeconomic effects. Environmental Evidence, 5, 21.
117. ROONEY, A. A., BOYLES, A. L., WOLFE, M. S., BUCHER, J. R. & THAYER, K. A. 2014. Systematic review and evidence integration for literature-based environmental health science assessments. *Environmental Health Perspectives,*122**,** 711-8.
118. ROSENTHAL, R., FODE, K. 1963. The effect of experimenter bias on the performance of the albino-rat. *Behavioural Science*. 8,183-9.
119. ROTHSTEIN, H. R., SUTTON, A. J. & BORENSTEIN, M. 2005. Chapter 1. Publication bias in meta-analysis. *In:*Rothstein, H. R., Sutton, A. J. & Borenstein, M. (Eds.) *Publication bias in meta-analysis—prevention, assessment and adjustments.* Chichester, UK : John Wiley & Sons, Ltd.
120. SAGOO, G. S., LITTLE, J. & HIGGINS, J. P. T. 2009. Systematic Reviews of Genetic Association Studies. *PLoS Medicine,* 6**,** e1000028.
121. SALEH, A. A., RATAJESKI, M. A. & BERTOLET, M. 2014. Grey literature searching for health sciences systematic reviews: a prospective study of time spent and resources utilised. *Evidence Based Library and Information Practice,* 9**,**28-50.
122. SARGEANT, J. M. & O'CONNOR, A. M. 2014. Conducting systematic reviews of intervention questions ii: relevance screening, data extraction, assessing risk of bias, presenting the results and interpreting the findings. *Zoonoses and Public Health,* 61**,** 39-51.

123. SAYERS, A. 2007. Tips and tricks in performing a systematic review. *British Journal of General Practice,* 57**,** 759.
124. SCARGLE, J. D. 2000. Publication Bias: The "File-Drawer" Problem in Scientific Inference. *Journal of Scientific Exploration* 14: 91–106.
125. SCHAFER, J. L. 1997. *Analysis of incomplete multivariate data. Chapman & Hall/CRC Monographs on Statistics & Applied*. CRC Press.
126. SCHINDLER, S., LIVOREIL, B., PINTO, I. S., ARAUJO, R. M., ZULKA, K. P., PULLIN, A. S., SANTAMARIA, L., KROPIK, M., FERNANDEZ-MENDEZ, P. & WRBKA, T. 2016. The network BiodiversityKnowledge in practice: insights from three trial assessments. *Biodiversity Conservation,* 25**,** 1301-1318.
127. SCHULZ, K., CHALMERS, I., HAYES, R. & ALTMAN, D. 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*, 273, 408-12.
128. SHARP, S. 1998. Meta-analysis regression: statistics, biostatistics, and epidemiology. *Stata Technical Bulletin* 42: 16-22.
129. SMART, J. M. & BURLING, D. 2001. Radiology and the Internet: a systematic review of patient information resources. *Clinical Radiology,* 56**,** 867-870.
130. SMITH, R. K., PULLIN, A. S., STEWART, G. B. & SUTHERLAND, W. J. 2010. Is predator control and effective strategy for enhancing bird populations? CEE review 08-001 (SR38). Collaboration for Environmental Evidence: www.environmentalevidence.org/SR38.html.
131. SOCIAL SCIENCE RESEARCH UNIT. 2016. Eppi Reviewer 4. London: Social Science Research Unit, Institute of Education, University of London. https://eppi.ioe.ac.uk/cms/Default.aspx?alias=eppi.ioe.ac.uk/cms/er4
132. SÖDERSTRÖM, B., HEDLUND, K., JACKSON, L. E., KÄTTERER, T., LUGATO, E., THOMSEN, I. K. & JØRGENSEN, H. B. 2014. What are the effects of agricultural management on soil organic carbon (SOC) stocks? *Environmental Evidence,*3**,** 2.
133. SONG, F., PAREKH, S., HOOPER, L., LOKE, Y. K., RYDER, J., SUTTON, A. J., HING, C., KWOK, C. S., PANG, C. & HARVEY, I. 2010. Dissemination and publication of research findings: an updated review of related biases. *Health Technology Assessment,* 14.
134. STEVENS, A., & MILNE, R. 1997. "The effectiveness revolution and public health". In: Scally, G. (Ed) Progress in Public Health, pp. 197-225. Royal Society of Medicine Press, London.
135. STEWART, G., COLES, C., & PULLIN, A. 2015. Applying evidence-based practice in conservation management: Lessons from the first systematic review and dissemination projects. *Biological Conservation*.126, 270-8.
136. STEWART, R. & LIABO, K. 2012. Involvement in research without compromising research quality. *Journal of Health Services Research and Policy* 17, 248-251.
137. THOMPSON, S. G. 1994. Systematic Review: Why sources of heterogeneity in meta-analysis should be investigated. *BMJ,* 309**,** 1351.
138. TRAMÈR, M. R., REYNOLDS, D. J. M., MOORE, R. A. & MCQUAY, H. J. 1997. Impact of covert duplicate publication on meta-analysis: A case study. *British Medical Journal,* 315**,** 635-640.

139. VON ELM, E., POGLIA, G., WALDER, B. & TRAMÈR, M. R. 2004. Different patterns of duplicate publication: An analysis of articles used in systematic reviews. *JAMA,* 291**,** 974-980.
140. VON ELM E., TRAMÈR, M. R, JÜNI, P. & EGGER, M. Does duplicate publication of trials introduce bias in systematic reviews? A systematic review [abstract]. In: 11th Cochrane Colloquium: 2003 Oct 26-31; Barcelona, Spain.
141. WOOD, L., EGGER, M., GLUUD, L., SCHULZ, K., JÜNI, P., ALTMAN, D., et al. 2008. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ.*, 336, 601-605.
142. WORTMAN, P. 1994. Judging research quality. In: Cooper H. & Hedges L. (Eds), The Handbook of Research Synthesis. New York: Russell Sage Foundation. pp. 97-109
143. ZHANG, L., SAMPSON, M. & MCGOWAN, J. 2006. Reporting the role of expert searcher in Cochrane Reviews. *Evidence Based Library and Information Practice,* 1**,** 3-16.